# The Causal Argument for Physicalism

David Yates
King's College London
Doctoral Dissertation 2005

Word count: 95,495 (pp.4-244)

Acknowledgements
I gratefully acknowledge the financial support of the AHRC while undertaking full-time research on this project, from September 2002 to September 2004.
I am extremely indebted to Professor David Papineau of King's College London, who supervised both my full-time research during the above period, and my subsequent writing-up.

*Abstract*

Physicalism can be thought of as the view that the mental is "nothing over and above" the physical. I defend a formulation of this view based on supervenience. Physicalism may be supported in two ways: either by providing an explanatory account of the mind in physical terms, or by philosophical argument. Since we have only a rudimentary scientific understanding of the mind, physicalism needs argument. The most promising such argument is the causal argument, which may be summarised thus: (i) mental properties have physical effects; (ii) physics is causally complete (all physical effects have physical causes); (iii) effects are not generally overdetermined; so (iv) mental properties are physical. Of these premises, (i) relies on common-sense, (ii) relies on empirical support, and (iii) is *a priori*. I consider the merits of this argument by articulating two kinds of mental property emergence, 'weak' and 'strong', both of which are incompatible with physicalism. I show that the premises of the causal argument are compatible with weak emergence, and that the argument is therefore not deductively valid. The causal argument establishes that one of physicalism or weak emergence is true. However, weak emergence is problematic in ways that physicalism is not. If these problems are serious, then physicalism is to be preferred on other grounds, such as theoretical elegance and simplicity. However, I proceed to show that the *soundness* of the argument is questionable, as premise (ii) is unsupported by the available evidence. Strong emergence is inconsistent with (ii), so evidence for (ii) must (on reasonable assumption) be evidence against strong emergence. But all currently available evidence is consistent with strong emergence, and so this evidence does not support (ii). Future evidence might, but I argue that such evidence would need to involve the kind of scientific account of mind the lack of which motivates the causal argument in the first place. A well-supported causal argument, by the nature of the justification necessary for (ii), is otiose.

## CONTENTS

# Introduction

The purpose of this work is to assess the merits of the causal argument for physicalism about the mind. Physicalism about the mind is the claim that the mental is nothing over and above the physical. The causal argument for physicalism is a general argument scheme, which, if successful, is capable of establishing physicalism about any domain of causes that has physical effects. The argument is deceptively simple, and is based on three key premises. Two of these premises will be common to all causal arguments; they are the completeness of physics, and the no-overdetermination rule. The completeness of physics states that all physical effects have sufficient physical causes. The no-overdetermination rule states that in general, events do not have more than one distinct sufficient cause. The third premise depends on the domain of causes for which physicalism is to be established, and is the claim that those causes have physical effects. Since my concern is with physicalism about the mind, the third premise that concerns me is the claim that mental causes have physical effects. By completeness, we know the physical effects of mental causes have sufficient physical causes. But by the no-overdetermination rule, we know that those effects do not have more than one distinct sufficient cause. The conclusion is that the mental causes of physical effects are not distinct from their physical causes. For the purposes of this work, I will take the causes and effects in question to be events. The causal argument, as I will conceive it, purports to establish physicalism about the mind by establishing that mental events are not distinct from physical events.

My overall aim in what follows is to show that this argument is neither sound nor valid. It is invalid because there are non-physicalist positions that are consistent with its premises, properly interpreted. It is unsound because the completeness of physics lacks crucial empirical support – there is nothing in the state of current science to suggest that all physical effects have sufficient physical causes. The invalidity of the argument, I will argue, is not fatal, for the non-physicalist positions in question are dubious on independent grounds. Its unsoundness, on the other hand, is fatal. The reason we need the causal argument for physicalism about the mind in the first place,

is that we lack a reductive account of the mind in physical terms, which would support physicalism empirically. I will argue that the only way to provide empirical support for the completeness of physics, is through scientific reductions of the very domains about which the causal argument is supposed to establish physicalism. This being the case, a sound causal argument (should we ever possess one) will be an argument for something we already know.

My arguments for the crucial claims of this work – viz., that the causal argument is neither sound nor valid – will depend on the coherence (but not the truth) of emergentism about mental properties. Correspondingly, the problems I raise for the argument will concern its ability to establish physicalism about mental properties. Such problems need only trouble physicalists who are realistic about properties. I offer no grounds to doubt that physical substance monism is true, and in fact endorse monism myself. Those physicalists who believe, for instance, that a 'property' is just a class of particulars that we (for whatever reason) group together under a concept, need not lie awake at nights wondering whether I am right. I take it as my starting point that properties are real, distinct from particulars, and that physicalism makes claims about both categories. It will not be sufficient, then, for the causal argument to establish that certain relations (such as identity or composition) hold between mental and physical event tokens – it must also establish that appropriate relations hold between the mental and physical properties of those events.

Here is a brief summary of each chapter:

Chapter 1

Since it would be of little value to attempt to determine whether or not physicalism is true without first determining what it is, I address in this chapter the question of how physicalism ought to be defined. 1.1-1.3 form a defence of global supervenience as a formulation of physicalism. 1.4 relates the formulation of physicalism to the relations that obtain between token mental and physical events if physicalism is true. By relating supervenience to sufficiency relations between token events, I show how if

the causal argument can establish that the right kind of sufficiency relations hold between mental and physical events, it will establish physicalism.

Chapter 2

The purpose of this chapter is to motivate the causal argument. I argue that the reason we need an argument for physicalism about the mind is that we lack a reductive account of the mind in physical terms. I give an account of Kim's functional reduction, and show how it is ideally suited to establishing the supervenience thesis which, I argued in chapter 1, defines physicalism. 2.1-2.2 describe functional reduction, and contrast it with classical, Nagelian reduction. 2.3-2.4 argue (against Kim) that functional reductions establish supervenience rather than type identities. 2.5 speculates as to where we are at present with respect to a functional reduction of mind, thereby showing why the causal argument is needed.

Chapter 3

In this chapter I look in detail at the premises of the causal argument, and give a precise formulation of it. 3.1 is concerned with the efficacy of the mental. Since I do not doubt that mental events cause physical events, I have little to say about this matter that is new. 3.2 is concerned with the completeness of physics. I give a formulation of an empirically based argument for completeness, due to Papineau, which relies on the success of physiology in explaining bodily movements. 3.3 examines the no-overdetermination rule, linking it to the impossibility of widespread coincidence. I examine Kim's causal exclusion principle, arguing that the sufficiency relation defined in 1.4 is enough to render the co-occurrence of co-occurrent events non-coincidental. 3.4 gives a detailed run-through of the argument, and shows how it establishes supervenience via sufficiency relations between token mental and physical causes. I suggest that weaker sufficiency relations than that required to establish physicalism will suffice to dispel the coincidences ruled out by the non-overdetermination rule. Through this I suggest that there may be non-physicalist forms of supervenience that are consistent with the premises of the causal argument.

Chapter 4

In this chapter I respond, on behalf of the causal argument, to a charge of equivocation raised by Sturgeon. If the argument equivocates, then it will not establish any form of supervenience, physicalist or not. I show that the argument does not equivocate. 4.1 details Sturgeon's argument that the causal argument needs causal 'transmission principles' in order for it to generate 'causal competition' between mental and physical causes. 4.2 examines and rejects Sturgeon's argument against these transmission principles. In 4.3, I suggest arguments of my own as to why the principles might fail. And 4.4 shows that the causal argument does not need the principles in the first place, because contrary to popular belief, it does not require that mental and physical causes be in competition for the same effects in order to work.

Chapter 5

Here I offer a detailed study of Kim's much-discussed causal exclusion argument. This argument, if cogent, can be marshalled to establish type identity physicalism. My overall aim in this chapter is to show that the exclusion argument rests on a false theory of causation. 5.1 examines two versions of the exclusion argument, showing how each rests on an unsupported, stronger version of the no-overdetermination rule than the one justified by the impossibility of widespread coincidence. In 5.2, I describe a theory of causation, based on causal work, to which Kim is plausibly committed, and which entails the unsupported premise. 5.3 gives a general formulation of the resulting causal exclusion problem, and 5.4 taxonomises possible responses to this problem. In 5.5, I argue that by far the most plausible response to the exclusion problem is to reject the theory of causation upon which it is based. This theory, I will argue, is dubious on independent grounds. There is no causal exclusion problem; the causal exclusion argument does not work.

Chapter 6

In this chapter I appeal to emergence in order to define a non-physicalist position, which I term 'weak emergence', that is consistent with the premises of the causal argument. 6.1-6.3 define two forms of emergence, 'strong' and 'weak', by combining

a general metaphysical claim about emergent properties with two distinct claims about their causal novelty. In 6.1, I describe the general metaphysical claim. 6.2 offers a general causal theory of novelty and redundancy, according to which there are two kinds of novelty, and shows how this enables us to resist an argument, due to Kim, to the effect that all supervenient properties are redundant given the completeness of physics. 6.3 defines weak and strong emergence based on the metaphysics of 6.1 and the two varieties of novelty defined in 6.2. 6.3 concludes with a discussion of why weak emergence renders the causal argument invalid. 6.4-6.5 offer two related arguments against weak emergence. 6.4 argues that weakly emergent mental properties would be redundant. 6.5 argues that weakly emergent properties in general suffer from the so-called 'miraculous coincidence problem'. This chapter concludes that although the causal argument is not deductively valid in its own right, this problem can be remedied by combining it with additional arguments.

Chapter 7

Strong emergence as defined in chapter 6 is inconsistent with the completeness of physics. But then evidence for completeness ought to be evidence against strong emergence. In this chapter I show that there is no current evidence against strong emergence. 7.1 examines the putative evidence for completeness, and 7.2 considers and rejects this evidence. The argument from physiology is refuted. 7.3 discusses and rejects Kim's supervenience argument against emergence, and through this shows what good evidence for the completeness of physics would look like. 7.4 argues that what this good evidence amounts to is the very functional reduction of mind the absence of which motivates the causal argument in the first place. The only way the completeness of physics can be empirically supported is if we already have enough evidence to infer physicalism without additional argument. At worst, the causal argument is unsound; at best, it is otiose.

# 1. What is Physicalism?

A physicalist must maintain that in some sense, "everything is physical"; but what's the sense? Some things, like quarks and charge, are transparently physical; others, like mentality and life, are not. The desiderata for a successful definition of physicalism would seem, *inter alia*, to be to capture the thought that the non-transparently physical things are (i) determined by, (ii) dependent upon and (iii) nothing over and above, the transparently physical things. I follow the general consensus in taking the concept of supervenience to be the most promising route to such a definition, but it should be noted that the purpose of this chapter is not to give a complete defence of any supervenience thesis with regard to the above constraints. Rather, its purpose is to set up the groundwork for the discussion to follow, and as such it will leave a great deal to be said. Since (as is the norm) I will be running the causal argument in terms of token events, I will first discuss supervenience based formulations of physicalism relating sets of mental and physical properties, and then show how this extends to the relations that obtain between mental and physical event tokens if physicalism is true.

In 1.1, I discuss and compare what I take to be the two most popular candidate definitions in the literature, namely 'global' and 'strong' supervenience. In 1.2, I argue that strong supervenience collapses into global supervenience if stripped of certain implausible commitments. The global supervenience thesis I endorse may well be too weak for many physicalists – indeed, there are arguments to the effect that it is insufficient to meet any of (i)-(iii) above. I will consider these arguments in 1.3, and defend global supervenience against them. My defence is somewhat tentative, which would be an issue if my overall aim were to defend the causal argument. However, most are in agreement that the global supervenience thesis I favour is a *necessary* condition for physicalism, and since my overall aim is to highlight certain problems with the causal argument, a necessary condition is all I need. If the argument is unable to establish a plausibly necessary condition for physicalism, then *a fortiori* it will fail to establish physicalism. In 1.4, I give an account of how events can be non-causally *sufficient* for other events, and give a definition of the appropriate sufficiency relation.

By appealing to this account, I will show that (i) on reasonable assumption, we can infer supervenience from sufficiency; and (ii) if it is to establish supervenience *physicalism*, the causal argument must establish at least that sufficiency of a particular strength (which I will term 'physical sufficiency') holds between physical and mental events.

### 1.1. Global and Strong Supervenience

The central intuition behind the supervenience of A-properties on B-properties, as is by now familiar, is that there can be no difference in A-properties without some difference in B-properties. The idea was suggested by Hare in the context of discussing naturalism about evaluative properties. Hare's by now familiar example is that there could not be two pictures, alike in all respects, one of which is praiseworthy as a work of art, the other of which is not.[1] Of course, as Hare proceeds to argue, unless the appropriate respects of similarity are specified without reference to praiseworthiness, the supervenience of artistic merit on these properties will be vacuous. The natural way to think about the matter is that two works of art alike in all intrinsic, physical respects could not differ as to their artistic merit. The determination in question is *asymmetric*, in that there could be no difference in the aesthetic properties without some physical difference, but that there are some physical differences that make no difference to artistic value. Since then, largely thanks to work by Kim, supervenience has come to prominence as a way of formulating physicalism.[2] A plausible initial attempt is the following 'global supervenience' thesis, defining physicalism for the actual world $w_a$:

P1. Physicalism is true at $w_a$ iff any world $w_x$ that is a physical duplicate of $w_a$ is a duplicate of $w_a$ *simpliciter*.

[1] See Hare [1963] 5.1.
[2] See for instance Kim [1990].

Roughly put, P1 states that that physicalism is true at the actual world if and only if any world that shares the actual world's distribution of physical properties and particulars, shares the actual world's distribution of all other properties and particulars as well. It is not, of course, mandatory for a global supervenience thesis to define physicalism as a claim about the actual world only – for instance, we could make it a claim about all (at least physically possible) worlds, viz. 'physicalism is true iff any two worlds that are physical duplicates are duplicates *simpliciter*.' Now although there seems to be *something* clearly right about P1, it will not suffice as a definition of physicalism. This is because it suffers from what Lewis has termed the 'problem of extras' – P1 rules out a world that is physically indistinguishable from the actual world, but differs from it in containing some extra "epiphenomenal ectoplasm", say. This (by definition non-physical and undetectable) stuff just floats around minding its own business, passing through the physical stuff completely unnoticed. There is no pressure to regard an ectoplasm world as one at which physicalism is *true*, but surely the truth of physicalism ought not to entail that such worlds are *impossible*. The trouble, then, is that physicalism should leave open the possibility of 'extras', yet P1 does not.[3] Jackson suggests the following revision of the thesis to accommodate ectoplasm worlds:

P2. Physicalism is true at $w_a$ iff any world $w_x$ that is a minimal physical duplicate of $w_a$ is a duplicate of $w_a$ *simpliciter*.[4]

The intuitive idea P2 strives to capture is that if physicalism is true, then any world that is a physical duplicate of the actual world *but contains nothing else*, is a duplicate in all other respects as well. The 'contains nothing else' – that is, the 'minimal' aspect of duplication – excludes ectoplasm worlds from the antecedent of the conditional on the right of the equivalence. The truth of physicalism, on P2, no longer entails that such worlds are impossible, for they are by hypothesis not minimal physical

[3] Lewis [1983] p.35.
[4] Jackson [1993]. In fairness to Jackson, it should be noted that this point that he proposes P2 only as a *necessary* condition on the truth of physicalism. I give further consideration below (especially in 1.3) to the question whether P2 is also sufficient – my view is that it is, but this is a highly contentious matter.

duplicates of *any* world. Further, P2 has the virtue that physicalism comes out *false* at such worlds, considered as actual – for minimal physical duplicates of them will lack the ectoplasm, and so fail to be duplicates *simpliciter*. Here it is the notion of a 'minimal physical duplicate' that is doing the work. Now in both P1 and P2, the notion of duplication employed is an imprecise, intuitive concept, about which clearly something more needs to be said; I will now attempt to say some of it.

Lewis defines duplication for *individuals* in terms of shared intrinsic properties. However, as Lewis points out, it is natural to understand intrinsic properties as those cannot differ between duplicates.[5] As a means of breaking out of this definitional circle, Lewis and Langton propose an alternative definition of intrinsic. A lonely object is the sole occupant of the world it occupies; P is an intrinsic property just in case (i) there is a lonely P; (ii) there is a lonely non-P; (iii) there is a non-lonely P; (iv) there is a non-lonely non-P.[6] If P is an intrinsic property of *x*, then *x* would possess P whether or not there were any other objects in the world that *x* occupies. Take any wholly distinct object you like out of *x*'s world, put any such object you like in – you will not thereby affect *x*'s intrinsic properties. So we might try the following as a definition of minimal physical duplication for *worlds*: $w_1$ is a minimal physical duplicate of $w_2$ just in case $w_1$ and $w_2$ have all the same intrinsic physical properties. But clearly this proposal will not work if 'intrinsic' is defined in terms of possible worlds. The reason is that it makes no literal sense to classify possible worlds as lonely or otherwise – in order for a possible world to be lonely, it would have to occupy another possible world at which there was nothing else. We must therefore explain world-level duplication by reference to individuals. Suppose duplicate individuals to be those that share all the same intrinsic properties. One point is immediately clear: if you want to create a duplicate $w_2$ of $w_1$, it is not sufficient to

[5] In his [1983] pp.25-33. In that paper he proceeds to break out of the circle by defining duplication instead in terms of shared perfectly natural properties. The details of this account are well outside the scope of the present work.

[6] This is supposed to capture the idea that if a property is intrinsic, then *whether or not* an object has it is independent of *whether or not* there is anything else. There are certain complications arising from disjunctive properties, but they are beyond the scope of the present work. See Lewis and Langton [1998].

populate $w_2$ with all and only individuals that are duplicates of those at $w_1$. First, this will leave out any laws of nature that hold at $w_1$; we must therefore include these by stipulation. Second, it will leave out a great many relational properties; $w_2$ will not be a duplicate of $w_1$ if its individuals stand in different spatiotemporal relations, for instance. I tentatively propose that we can define duplication for worlds if we appeal to (i) duplication for individuals, as defined above; (ii) the laws of nature, and (iii) relations between the duplicate individuals. Let duplicates be individuals that share their intrinsic properties, and let us define intrinsic according to the model proposed by Lewis and Langton. I propose that:

> $w_2$ is a duplicate *simpliciter* of $w_1$ just in case (i) for every individual $x$ at $w_1$ there is exactly one individual $y$ at $w_2$ such that $y$ is a duplicate of $x$; (ii) every individual at $w_2$ is a duplicate of some individual at $w_1$; (iii) for any n-adic relation R and ordered n-tuple $x_1,\ldots,x_n$ at $w_1$ such that $R(x_1,\ldots,x_n)$, there is an ordered n-tuple $y_1,\ldots,y_n$ at $w_2$ such that for any $i$, the $i$th member of $y_1,\ldots,y_n$ is a duplicate of the $i$th member of $x_1,\ldots,x_n$ and $R(y_1,\ldots,y_n)$; (iv) any law that holds at $w_1$ or $w_2$ holds just in case it holds at $w_1$ and $w_2$.

Now if my proposal is successful, then it defines duplication *simpliciter* for worlds in terms of the intrinsic and relational properties of individuals, and the laws of nature. How are we to define *minimal physical* duplication? We might try modifying the above definition of duplication *simpliciter* by inserting 'physical' before occurrences of 'individual', 'duplicate' and 'relation', and replacing 'law' with 'law of physics'. Take 'physical individual' as a primitive – physical individuals are spatiotemporally located, have physical properties like mass, energy, charge, are governed by physical laws, and so on. (Physical individuals so construed may, of course, possess *non*-physical properties as well.) Physical duplicates will be physical individuals that do not differ in their intrinsic physical properties. So far so good; unfortunately, it is less than clear how we are to understand 'physical relation' in this context. In the general case, we can not understand physical relations as those that hold between particulars with physical intrinsic properties – for one such particular might more beloved of God

than another. We might try construing physical relations as those somehow *determined* by intrinsic physical properties and laws, but determined *how*? Clearly we cannot appeal to any supervenience thesis for physical relations that invokes the notion of a minimal physical duplicate, for then the definition takes us in a circle back to the *definiendum*. Fortunately, there is a way out. Imagine we are in the business of building worlds. We want to make it so that causal relations, particles exerting forces on each other, objects being in thermal equilibrium, and so on, are all fixed. My stipulation at this point is that all we would have to do as world builders is fix the intrinsic physical properties of everything, fix the laws of physics, and fix the spatiotemporal location of each object. I take as a further primitive that the laws of physics *govern* the behaviour of individuals, the forces they exert on each other, that things are in thermal equilibrium, *according to their intrinsic physical properties and positions in spacetime*. Now we have the following definition of minimal physical duplication:

> $w_2$ is a minimal physical duplicate of $w_1$ just in case (i) for every physical individual $x$ at $w_1$ there is exactly one physical individual $y$ at $w_2$ such that $y$ is a physical duplicate of $x$; (ii) every individual at $w_2$ is a physical duplicate of some physical individual at $w_1$; (iii) for any n-adic spatiotemporal relation R and ordered n-tuple $x_1,\ldots,x_n$ of physical individuals at $w_1$ such that $R(x_1,\ldots,x_n)$, there is an ordered n-tuple $y_1,\ldots,y_n$ at $w_2$ such that for any $i$, the $i$th member of $y_1,\ldots,y_n$ is a physical duplicate of the $i$th member of $x_1,\ldots,x_n$ and $R(y_1,\ldots,y_n)$; (iv) any law that holds at $w_2$ is true in all physically possible worlds; (v) any intrinsic property of any individual $x$ at $w_2$ is instantiated by any physical duplciate of $x$ at any physically possible world.[7]

A few notes are in order. *First*, I have included 'of physical individuals' after 'ordered n-tuple $x_1,\ldots,x_n$' in (iii) to allow for the possibility that any non-physical individuals at

[7] Note that if a definition of physical duplication – rather than *minimal* physical duplication – for worlds is wanted, then we plausibly get one by removing (ii) and (v) from the above definition, and rewriting (iv) as simply 'any law of physics that holds at $w_1$ holds at $w_2$'. Hereinafter, where I refer to physical duplicate worlds, I have in mind a definition of the form just mooted.

$w_1$ might stand in spatiotemporal relations; it is very important not to build physicalism into the definition of minimal physical duplication. It is not necessary to repeat this stipulation for $y_1,\ldots,y_n$, for we know by (ii) that all the $w_2$ individuals are physical. *Second*, as to (iv), a previous formulation read 'any law of physics that holds at $w_1$ or $w_2$ holds just in case it holds at $w_1$ and $w_2$'. But this would allow $w_2$ to be a minimal physical duplicate of $w_1$ and yet there be *non*-physical laws that hold at $w_2$ but not at $w_1$, which is clearly unacceptable. I could have said 'the only laws that hold at $w_2$ are laws of physics' but then what of psychological, geological, and chemical laws? Again, appeals to supervenient laws are to be avoided, on pain of circularity – we can not therefore write (iv) as 'the only laws that hold at $w_2$ are laws of physics and any other laws supervenient on them.' The appeal to quantification over physically possible worlds is supposed to avoid these pitfalls. The only laws common to all physically possible worlds will, of course, be either physical laws or consequences of physical laws. But appealing to the *physical necessity* of this set of laws in order to define its members avoids any tacit appeals to supervenience.[8] *Third*, similar considerations to those given in support of (iv) motivate (v) – this condition is intended to rule out individuals at $w_2$ that are physical duplicates of $w_1$ individuals and yet possess, for instance, Cartesian souls. Clearly not *all* the properties of $w_2$ individuals will be physical; but once again, we cannot appeal to supervenience to specify the acceptable *non*-physical properties. The problem is exactly analogous to the problem with laws; so, I take it to be relatively clear, is my suggested solution.

Now if we plug this definition of minimal physical duplication into P2, we get the thesis that physicalism is true at $w_a$ just in case any world $w_x$ formed by adding one particular for every actual physical particular, and nothing else, endowing each one with all and only the actual intrinsic physical properties of the corresponding $w_a$ particular, fixing it so that all $w_a$ spatiotemporal relations hold between the $w_x$ duplicates of $w_a$ individuals, and fixing it so that all and only physically necessary

[8] I hold that 'physically possible' means 'consistent with the laws of physics', and 'physically necessary' means 'true in all physically possible worlds'. On this account the laws of physics are physically necessary, as are any other laws entailed, or otherwise determined, by physical laws. I am not entirely satisfied with this solution, but it is the best one I have to hand at the time of writing.

laws hold at $w_x$, duplicates $w_a$ *simpliciter*. Notice that given my stipulation that spatiotemporal relational properties take care of 'physical relations', we could also formulate condition (iii) in the definition of minimal physical duplication equivalently in terms of dyadic relations between individuals and reified spacetime points.

A few technical comments, in no particular order, before proceeding. *First*, '_is a minimal physical duplicate of_' is non-reflexive and non-symmetric. A world containing ectoplasm will not be a minimal physical duplicate of itself, as it contains non-physical stuff that won't survive duplication (nonreflexivity); and a minimal physical duplicate $w_2$ of an ectoplasm world $w_1$ will not contain the ectoplasm, and so, quite obviously, $w_1$ is not a minimal physical duplicate of $w_2$ (nonsymmetry). *Second*, including the laws of physics in the notion of a minimal physical duplicate has the consequence that we do not need to worry about restricting the modal 'any' in P2: for any possible world in which the actual laws of physics hold will *ispo facto* be a physically possible world. We can therefore leave the modality unrestricted. *Third*, it is not clear that a duplicate *simpliciter* of a world counts as a distinct possible world. But if indistinguishable possibilia are not distinct, then P2 merely claims that if physicalism is true, then any minimal physical duplicate of the actual world *just is* the actual world. I reply that this is unimportant. P2 is still a substantive claim, as there will be plenty of worlds such that minimal physical duplicates of them are *not* duplicates *simpliciter*; worlds, that is, such that there are minimal physical duplicates of them to which they are not identical. *Fourth*, P2 will fail to define physicalism if there exist at $w_a$ *non*physical entities that are metaphysically necessitated by physical ones. For if there are such entities, then minimal physical duplicate of $w_a$ will be duplicates *simpliciter*, despite the fact that physicalism is not true at $w_a$. I reply that there are no such entities – if there are ghosts following me around as a matter of necessity, then the necessity is not metaphysical. Nor is it physical. If ghosts obey laws that connect them to physical individuals at $w_a$, then those laws hold in addition to, and independently of, the laws of physics, and so will not hold at minimal physical duplicate worlds of $w_a$.

P2 looks promising, but runs into two problems, which are *prima facie* devastating. Both problems stem from the fact that P2 says nothing about worlds that differ from actuality. *First*, P2 is consistent with the existence of ghosts with minds at nearby worlds. Let $w_g$ be a world that would be a minimal physical duplicate of $w_a$ but for the existence of a happy ghost, which never makes any difference to anything else at the world. (I leave open for now whether it is merely *de facto* true that the ghost does not disturb anything, or due to its having no causal powers. I return to this point in 1.3.) P2 is exempt from application to $w_g$, which makes it consistent with the truth of physicalism at $w_a$ that there are very close possible worlds at which physicalism is false. It might be objected on this basis that P2 fails to adequately capture either the *dependency* of the mental on the physical, or (correspondingly) the fact that the mental is 'nothing over and above' the physical.[9] Call this the *dependency problem*. The happiness of the ghost at $w_g$ clearly does not depend on physical properties for its instantiation; further, *its* happiness *must* be 'something over and above' the physical. *Second*, there is the so-called *wayward hydrogen atom problem*, which is that P2 is consistent with the existence of a world $w_1$, which minimally duplicates $w_a$ in all respects save the position of a single hydrogen atom, and at which, for instance, nothing has any mental properties. Further, the definition also fails to rule out physically identical individuals within the actual world, one of which possesses mental properties, the other of which does not.[10] Intuitively, these possibilities seem at odds with physicalism – for surely they show that global supervenience does not posit strong enough *connections* between mind and body to be physicalistically acceptable? If P2 is consistent with the dependency of mental properties on factors we know to be psychologically irrelevant, then how can it adequately capture the thought

[9] Hendel [2001], for instance, argues that global supervenience theses are insufficient to capture the thought that the mental is 'nothing but' the physical on the grounds that such theses allow that there are possible worlds where *non*-physical things have mental properties. On Hendel's account of 'nothing buttery' it does not matter whether such worlds are close to actuality or not – the instantiation of a mental property M by a non-physical being at *any* world is sufficient to render M something over and above the physical. I return to this point, and offer a limited rebuttal, at the end of 1.3 below.

[10] These points are due to Kim. See, for instance, his [1989b], [1990]. It is unclear whether the cases described *are* consistent with P2. Paull and Sider [1992] give (as one horn of a dilemma) an argument to the effect that the consistency is merely *prima facie*, as wayward atom worlds entail (on reasonable assumption) other possibilities that directly violate global supervenience. I discuss their argument in 1.3 below.

that given physicalism, physical properties *determine* mental properties? Putting these two objections together, we have reason to doubt that P2 can capture *any* of the desiderata suggested for a definition of physicalism at the outset. I think that both these objections can be defeated, and will say why I think this in 1.3.[11] For the moment, we simply note that these difficulties lead many philosophers to favour Kim's *strong supervenience* (hereinafter I omit 'Physicalism is true iff' for brevity):

P3. $\Box\forall x\forall \mathrm{M}\in M\{\mathrm{M}(x)\rightarrow\exists \mathrm{P}\in P[\mathrm{P}(x)\&\Box\forall y(\mathrm{P}(y)\rightarrow\mathrm{M}(y))]\}$ [12]

Here $M$ is the set of mental properties and $P$ the set of physical properties that subvene some mental property, and M and P are members of the respective sets. An attractive feature of this sort of formulation is that it makes it possible for us to define physicalism independently for different sets of properties. We might, for instance, hold that physicalism is true for intentional mental properties and yet false for phenomenal properties. By contrast, P1 and P2 are "all or nothing" definitions – on these formulations, physicalism will be either true of *all* properties instantiated at a world, or else false. The first modal operator in P3 is intended to convey the dependency of mental properties on physical properties, the second the determination of the former by the latter. We can translate it like this: necessarily, anything with a mental property M has a physical property P such that necessarily, if anything is P then it is M. The standard way of understanding the strengths of the modalities here is that the first operator expresses physical (sometimes nomological) necessity, the second metaphysical. Interpreting the first operator as physical necessity allows for the possibility *simpliciter* of spirit worlds, but disallows any non-physical beings with minds at physically (or nomologically) possible worlds.

P3 easily handles the dependency problem. The wide-scope modal operator in P3 means that the definition rules out the instantiation of any properties in $M$ at

[11] In particular, I will argue that the wayward atom 'problem' is actually a virtue of P2. Solving the dependency problem is much more difficult, and my response to it will be somewhat inconclusive.
[12] See Kim [1990] for details.

physically (or perhaps nomologically) possible worlds that lack the relevant properties in *P*. This is dependency with a vengeance – there is no individual, at *any* physically possible world, with mental properties and no physical properties. This clearly rules out *a fortiori* the troublesome ghosts that pose the dependency problem for P2. What of the wayward hydrogen atom problem? Provided *P* is restricted to intrinsic properties, the hydrogen atom problem described above is not a problem for P3, as P3 entails that duplicate physical *individuals* can not differ mentally. Of course it follows *a fortiori* that minimal duplicate *worlds* can not differ mentally either, but the converse entailment is clearly untrue, which is precisely why P2 is consistent with physical duplicate individuals (within or across worlds) that differ mentally in virtue of some small difference in the apparently insignificant physical properties of those worlds. For the same reason, P3 is also inconsistent with the existence of two physical duplicates *within* a world that differ mentally. However, Paull and Sider argue that the Hydrogen atom problem resurfaces in that P3 says nothing about *individuals* that differ in small but psychologically insignificant ways.[13] The wayward atom problem, they claim, can be internalised – for P3 is consistent with a physical 'almost-duplicate' of George Bush (who would be a physical duplicate save for one wayward atom in his brain) who lacks all of actual Bush's mental properties. This is because P3 "…only implies that an object must have all the wonderful mental properties that Bush actually has if the object shares *all* his physical properties".[14] This seems right – there is nothing in P3 that will ensure that individuals that differ in minute *P*-respects will differ (if at all) in correspondingly minute *M*-respects.

If Paull and Sider are correct, then strong and global supervenience stand or fall together in the face of their respective wayward atom problems. I note in passing that I am unconvinced that this is so. In particular, it is unclear to me that no restriction on *P* will rule out the case described. For instance, supposing *P* is allowed to contain only complex neural properties – the kind of properties that our best current theories of mind tell us do in fact determine our mental properties. With this restriction, two

[13] See their [1992] pp.841-2.
[14] Paull and Sider [1992] p.842.

individuals might well fail to be physical duplicates and yet still be *P*-duplicates – for the properties in *P* may be insensitive to fine-grained physical differences. Changing the position of a single atom in Bush's brain, it seems clear, will not change his neural properties. I will not concern myself with this matter further here, for the putative restrictions on *P* are implausible. So too is the wide-scope modal operator. For these reasons – or so I will argue in the next section – P3 has no advantage over P2 with either the dependency or wayward atom problems.

There are two central problems with P3 that push it towards a position more-or-less equivalent to P2. *The first* of these problems I will call the *problem of relational properties*. The problem is that restricting the set *P* to intrinsic properties is unjustifiable, for reasons discussed below. But unless *P* is restricted to intrinsic properties, P3 is really a *global* supervenience thesis (P4, defined below) not dissimilar (though also not equivalent) to P2.[15] The non-equivalence of P2 and P4 consists in (i) the fact that P4 rules out certain possibilities (to be described) not ruled out by P2, and (ii) that fact that P4 defines physicalism for mental properties, whereas P2 just defines physicalism. The problem of relational properties is a corollary of the wayward atom problem; P3 has the former precisely because it does not have the latter. Similarly, *mutatis mutandis*, for P2. *The second* problem I will call the *modality problem*. The problem is that the wide-scope modal operators in P3 and P4 have undesirable consequences that warrant their removal. But its removal from P4 yields a weaker global supervenience thesis (P6, defined below), which is more-or-less equivalent to P2. The non-equivalence of P6 and P2 consists solely in the fact that P6 defines physicalism for *mental* properties, whereas P2 defines *physicalism*. Correspondingly, if we broaden P6 so as to define physicalism simpliciter, it is equivalent to P2. The modality problem is a corollary of the dependency problem; P3 has the former precisely because it does not have the latter. Similarly, *mutatis mutandis*, for P2. In the next section I will explain why the relational properties and

[15] Horgan argues for a similar point in his [1993], but claims that strong and global supervenience theses are equivalent if *P* is unrestricted. I take this claim to be false due to the presence of the extra wide-scope modal operator in P3. I will justify this point in 1.2.

modality problems are serious enough to warrant rejection of P3. In 1.3 I will explain why the wayward atom and dependency problems are not serious enough to warrant rejection of P2.

## 1.2. Two problems for strong supervenience

1. The problem of relational properties

If we include only *intrinsic* properties in *P*, P3 rules out, for instance, various theories of content that claim that mental properties are determined by environmental factors. Such theories are numerous, but examples include Burge's conventionalist account of conceptual content, Putnam's argument that the semantics of natural kind terms are environmentally determined, and the various naturalization projects that hold contents to be determined by historical and/or causal factors. I will not go into any great detail, as the point I wish to make is quite general. Burge [1979] endorses the view that conceptual content is determined by correct linguistic practice, which in turn is socially determined. Even if I intend my use of the term 'arthritis' in 'I have arthritis in my thigh' to mean *cramp*, the content of my utterance is that I have *arthritis*, not cramp, in my thigh. But assuming that the content of my belief that I have arthritis in my thigh is the same as the content of the proposition, it follows (given that intentional states are partially typed by their contents) that my beliefs depend, *inter alia*, on socio-linguistic facts. If this theory is correct, then *P* must include these facts. According to Kripke's [1980] causal theory of naming, names do not have Fregean 'senses'; rather, they have reference only, which is determined by a causal process connecting an initial 'dubbing ceremony' to subsequent utterances. Putnam [1975c] extends Kripke's theory of names to natural kind terms, claiming that these too are ceremoniously named. Captain caveman says "I hereby name *this substance* 'water'" – and thereafter, 'water' means $H_2O$, and not 'watery, tasteless, colourless stuff'. On this account, the contents of our beliefs about natural kinds are determined by their deep physico-chemical structures, so *P* must include many (if not all) the physical properties of the believer's environment. Similar considerations apply to the causal covariance accounts of content suggested by Fodor [1987] and Dretske [1981]. If content is informational, and the information carried by a token depends on what

typically causes it, then mental content is again determined by environmental properties, which must then be included in *P*. Finally Millikan [1993] and Papineau [1993] endorse historical accounts of content based on teleological function. Content is informational on these accounts too, but mental tokens are said to carry information about only those environmental features they are *supposed* to covary with. On this account, *P* will have to include teleological properties. The crucial point I want to make here is that restricting *P* to intrinsic properties makes it incompatible with all the theories of content outlined. Clearly, it would be unwise for a physicalist to rule out such theories solely on account of being a *physicalist*.[16]

If, on the other hand, we include *relational* properties in *P*, then without some restriction on *which* of these properties can be included, the thesis (as I will argue) collapses into a form of global supervenience. And it seems, short of turning physicalism into a substantive *theory* of mind, that no such restriction can be supplied. For instance, what if we restrict the relational properties in *P* to synchronically

[16] As an aside, I should point out that I do not wish to endorse any such theory – like many others, I think that broad contents are insufficiently fine-grained to earn their psychological-explanatory keep. There are two central classes of problem case, and I follow Fodor [1994] in classifying them as 'twin-cases' and 'Frege-cases'. As to twin cases, suppose for the sake of argument that natural kind externalism is true. Twin Earth is a physical duplicate of the actual world save that it has XYZ (superficially indistinguishable from $H_2O$) instead of water. Suppose neither me nor my twin-Earth counterpart know anything about the chemical structure of the stuff in Earth and twin-Earth lakes, respectively. My twin's desire for some of what he refers to as 'water' is a desire for XYZ, not *water*; ditto for beliefs. And yet me and my twin apparently behave in just the same way because of our respective desires – we go to the kitchen, turn on the tap, hold a glass under it, and so on. Broad psychology appears to miss generalisations it ought to capture. As to Frege cases, suppose that Kripke's causal theory of names is true. But now (given that Cicero = Tully) the content p1 of Bob's belief that Cicero was put to death on Dec. 7, 43 B.C. is the same as the content p2 of the proposition that Tully was put to death on Dec. 7, 43 B.C. But Bob does not know that Cicero = Tully. Broad psychology thus struggles to explain certain behavioural facts: why Bob answers 'yes' when asked whether it is true that p1, but 'no' when asked whether it is true that p2; why Bob claims he does not know when Tully died; and so on. Broad psychology apparently gets the phenomena wrong. Fodor [1994] has a useful discussion of the relationship between broad content and the two problem cases mentioned. Fodor asks how, given that content is broad but computation narrow, we can reconcile the computational nature of psychological processes with the existence of content-based psychological laws. Twin cases and Frege cases are examples of how the narrow computational processes that cause behaviour, and broad-based psychological laws, can come apart. Fodor's answer on behalf of broad content based psychology is that such cases are sufficiently rare that the missed generalisations and predictive failures go unnoticed, and are as such unimportant. It is plausible that Fodor is right about twin-cases; but he is surely wrong about Frege cases. Individuals are frequently known by more than one name or description, which immediately raises the possibility of someone having beliefs about a given object under two or more 'modes of presentation', without the believer knowing that the different modes present the same thing.

instantiated properties, but it turns out that teleological theorists are right, and the determinants of content include *historical* factors? What if we restrict them to "surface features" of the subject's environment, but it turns out that natural kind externalists are right, and contents are determined by deep chemical structure? Restricting *P* in *any* way would seem to place undue *a priori* constraints on theories of mental representation. But without any restriction on *P* it seems we must insist that the property P that is sufficient for M must be a maximal structural property reflecting the total physical state of the world at which the individual has it. The only way to make sure all the *relevant* properties get into the supervenience base is to give up on trying to exclude the *irrelevant* ones, and let them all in. The very feature that (arguably) protects P3 from the wayward atom problem turns out to involve untenable commitments.[17] As in P2, we can appeal to the notion of a minimal physical duplicate to define this maximal set of subvenient properties.

The brand of physicalism we get from allowing unrestricted relational properties in P3 is similar, but as I said, not *equivalent to* P2. It will be helpful to express the thesis in terms of Lewis's counterpart theory;[18] let $W(x) = x$ is a possible world and $W^{\alpha}(x) = x$ is an α-possible world; let $I(x,y) = x$ is in possible world $y$, and $C(x,y) = x$ is a counterpart of $y$. For generality, let $P^{w-}(x,y) = x$ is a minimal physical duplicate of $y$. Then we can reformulate P3 in the following way:

P4. $\forall u \forall v \forall M \in M\{W^{\alpha}(v).I(u,v).M(u) \rightarrow \exists w \exists x[W(w).P^{w-}(w,v).C(x,u).I(x,w).$
$\forall y \forall z\{W(z).I(y,z).C(y,u).P^{w-}(y,v) \rightarrow M(y)\}]\}$

Thankfully, we may express this thesis in English as follows: 'If any individual $u$ in any α-possible world $v$ has a mental property M, then $u$ has a counterpart $x$ at a world $w$ that is a minimal physical duplicate of $v$, and any counterpart $y$ of $u$ in any possible

[17] In 1.3, I will suggest, correspondingly, that consistency with wayward atom worlds is a strength, not a weakness, of global supervenience.
[18] See Lewis [1968] and [1986a] for details. I will utilise counterpart theory quite a lot in what follows, as I find it heuristically invaluable. I do not endorse possible worlds realism – rather I hope that, pace Lewis, the relevant bits of the theory can be made to make sense with ersatz worlds of some kind. These matters are well beyond the scope of the present work.

world *z* that is a minimal physical duplicate of *v*, has M'. The quantification over *v* may be restricted by the predicate $W^{\alpha}$ if desired – we can let the wide-scope quantifier range over metaphysically, physically or nomologically possible worlds. I discuss such restrictions presently, when I come to the modality problem. While P3 claims that any individual with a mental property also possesses a physical property that is sufficient for its mental properties, P4 claims only that individuals with mental properties at a world have counterparts with the same mental properties at all minimal physical duplicates of that world. What is the advantage of this strategy? Well, by definition of counterparthood and $P^{w-}$, it follows that *x* and *y* will not only be physical duplicates of *u*, as counterparts across minimal physical duplicate worlds must be, but also possess all the same relational properties as well.[19] In this way, we succeed in making sure that we do not exclude any properties that might turn out to be relevant to the determination of mentality, while at the same time expressing the *dependency* claim that any individual at any possible world with mental properties must also have physical properties. This is because an individual with no physical properties fails to have counterparts at other worlds that are minimal physical duplicates of its own. This is also why the existential quantifiers are needed in P4 – without them P4 would fail to rule out worlds with mental but no physical properties, for minimal physical duplicates of such worlds fail to contain anything at all, making the antecedent of the second conditional false. Let us proceed to compare and contrast P4 and P2.

P4 clearly suffers from the wayward atom problem, as it is consistent with a world that *would* be a minimal physical duplicate of *w* but for the position of the wayward hydrogen atom, and which contains no mentality. It follows, of course, that P3 solves the wayward atom problem only if the set *P* is restricted to intrinsic properties, which, as I have argued, is a highly implausible restriction. P4 does not, however, have the dependency problem. The wide-scope quantifiers of P4, just as with the wide-scope

[19] Lewis, of course, defines counterparthood in terms of resemblance, and as such, I will certainly have counterparts that lack some of my mental properties. This does not, however, pose a problem for P4, which entails not that I have no such counterpart, but rather that I do not have such a counterpart in a world that is a minimal physical duplicate of the actual world. Since P4 is not concerned with worlds that differ from each other physically, it is quite consistent with defining counterparthood in Lewis's relatively weak terms.

modal operator in P3, means this problem does not arise. Any individual, at any α-possible world *w*, that has a mental property, has a counterpart at a world that is a minimal physical duplicate of *w*. But this, as I said above, means that any individual, at any α-possible world, that has a mental property, is a physical individual. P4, like P3, has dependency in spades. To see this, let $W^{\alpha}(x) = x$ is a physically possible world. Let $w_a$ be the actual world, and suppose that physicalism is true here. Once again let $w_g$ be a physically possible world that fails to minimally duplicate $w_a$ only in that there is also a happy ghost whose existence, we may suppose, is sufficient to render physicalism false at $w_g$. This scenario is consistent with P2, for as I take it is by now familiar, P2 says nothing about worlds that differ from the world for which physicalism is being defined. But according to P4, since the ghost has mental properties, there ought to be a minimal physical duplicate $w_g^-$ of $w_g$ containing a counterpart of the ghost. By definition of $P^{w-}$, the ghost at $w_g^-$ must be physical, for $w_g^-$ does not *contain* anything *non* physical. Now if the ghost at $w_g$ is *not* physical (which, by hypothesis, it is not), it is *false* that $P^{w-}(w_g^-,w_g)$, for in this case $w_g^-$ contains a physical individual with no counterpart at $w_g$. Therefore the ghost at $w_g$ must be physical, hence (contrary to stipulation) not the kind of ghost whose existence is inconsistent with physicalism. Ghosts are ruled out by P4 in that it quantifies over *all* physically possible worlds that contain individuals with mental properties. In sum, there is a modal difference between P2 and P4, in that the former holds that physicalism can be true at a world that has very close neighbouring worlds at which physicalism is *not* true, whereas the latter holds that physicalism, if true, is true of entire neighbourhoods of worlds – in our example, all the physically possible worlds. It is common in the literature to find philosophers insisting on just this latter sort of modal dependency of the mental on the physical; a moment's reflection, however, will show that this requirement is far too strong to be plausible. This leads us nicely into the second of our two problems.

## 2. The modality problem

The wide-scope quantification in P4 must be restricted somehow. This is because if we quantify over all possible worlds *simpliciter*, then the formulation rules out the

possibility *simpliciter* of 'spirit worlds' – for it would then claim (*inter alia*) that any individual in any possible world is such that if it has mental properties, then it has physical properties. Many, myself included, would find this far too strong. If physicalism is true, then there are certainly no ghosts at the actual world, and plausibly none at nearby worlds, but why should it be part of a physicalist's metaphysical commitments that ghosts are *impossible*? Our definition should not entail that there are no possible non-physical things that have mental properties. However, on reflection it is less than clear that *physical* necessity fares any better, for the simple reason that it is not obvious why ghosts should be physically impossible either. Let 'physically necessary' mean true in every physically possible world, and let 'physically possible' mean 'consistent with the laws of physics'. *Prima facie*, physical laws do not rule out the intervention of ghosts, Gods, or any other non-physical stuff in the causal workings of the world, since such laws do not *quantify* over such non-physical entities. Take a physical law of the form 'P events cause Q events'. An exception to such a law will presumably be the occurrence of a P event that is not followed by a Q event. But now suppose that, on some occasion, a ghost decides to intervene and cause a Q event. Unless its intervention prevents some other physical event from *preventing* this Q event, then its action does not violate physical law, as there is no physical law to the effect that ghosts never cause Q events! On the other hand, the ghost's intervention *will* violate the *causal completeness of physics*, if the Q event it causes lacks a physical cause. However, in the absence of an argument that the completeness of physics is physically *necessary* – and as we shall see in 3.2, it is no trivial matter to argue that it is even *true* – worlds containing ghosts and Gods and all sorts of other creepy things will still count among the physically possible worlds. But this means that the worlds at which physicalism is true will be a *subset* of the physically possible worlds, and so the wide-scope modality of P4 is too strong.

Can we further restrict quantification so as to make the definition plausible? It is difficult to see how. Consider *nomological* necessity. I understand this necessity to be truth in all possible worlds with the same *laws of nature* as the actual world. I distinguish between nomological and physical necessity because it is an open question

whether all laws of nature are, or are reducible to, laws of physics. If the world obeys 'extra' laws whose truth is independent of the truth of physical laws, then there will be fewer nomologically possible worlds than there are physically possible worlds, and nomological necessity correspondingly weaker than the physical variety. (Since the laws of physics in this case will be among the laws of nature, the set of physically possible worlds will be a superset of the nomologically possible ones.[20]) Now if the nomologically possible worlds are a subset of the physically possible ones, it is open to us to maintain that although ghostly interventions are physically possible, they are inconsistent with the extra laws that hold in the nomologically possible worlds. But how can that be so? These extra laws, although not determined by physical laws, will nonetheless be *natural* and amenable to empirical enquiry, which, I suppose, is exactly what ghosts are not! How *could* a natural law quantify over definitionally *super*natural entities?

One response at this stage is to adopt a richer notion of possibility. Instead of construing X-possibility as merely that which is not ruled out by X-laws, we might adopt a version of Armstrong's combinatorialism, in which (very roughly) a possible world is one that can be constructed by recombining the basic properties and particulars that exist at the actual world in any way you like.[21] Possible worlds so understood must of course be compatible with the actual laws of physics, but it must also be possible to construct them by recombining actual physical particulars and properties. This is not supposed to be an account of mere physical possibility, but rather of the *whole* space of possibilities. Any apparent possibilities that can't be made by recombining what's already here, are possible in name only, and must be reinterpreted as such. Now clearly you can't construct a ghost in this way, and so ghosts are impossible *simpliciter*. Their apparent possibility must then be re-

[20] This point is well made by Witmer [2001]. Witmer has this to say (pp.62-3): "By definition, every nomologically possible world is a physically possible world, but it is an open question whether every physically possible world is a nomologically possible world. It may be that there are laws not necessitated by the laws of physics, in which case there are physically possible worlds that are not nomologically possible." This point is of crucial importance, as we shall see in chapter 6.

[21] For the combinatorial theory see Armstrong [1986]. See Lewis [1992] and [1986b] for detailed discussion and criticism.

interpreted as something like (mere) epistemic possibility. Ghosts, while not ruled out *a priori* by the structure of our concepts, are no less impossible for all that. The trouble is, you can't construct new basic particulars out of old ones either, nor can new basic universals be produced by recombining old ones. And yet it seems perfectly possible that the world's basic ontology might have been otherwise, and not merely in the sense that it isn't *a priori* that it is the way it is.

We might appeal to combinatorialism as an account of physical possibility only, but it seems too strong even for that purpose. For instance, it seems clearly physically possible that there might, in addition to everything else that's here, have been one more quark in the world. But by definition you can't make the extra one by recombining the existing ones! If this objection is correct, then the set of combinatorially possible worlds is a subset of the physically possible worlds. These matters are well beyond the scope of this thesis, however; my view is that Armstrong's theory rules out far too many genuine possibilities to be plausible, but this is not the place to attempt a detailed justification. Suffice it to say that on a more liberal understanding of the space of possibilities, it looks very much like *any* restriction on the wide-scope quantifier in P4 (and correspondingly, any interpretation of the wide-scope operator in P3) will yield too strong a thesis. Problematically, the features that enable P3 and P4 to avoid the dependency problem, entail that certain genuine possibilities are not possible.

Let us try removing the wide-scope quantification over worlds altogether, making physicalism a claim about the actual world $w_a$.[22] Consider P5, which is a universal instantiation of P4 got by setting the value of the variable $v = w_a$:

[22] Clearly this move will bring strong supervenience much closer to the global variety. To some it will seem as though removing the wide-scope operator also removes a crucial component of P3 that expresses the dependency of the mental on the physical; the justification for its removal, after all, is precisely to allow for the physical possibility of *non*-physical things with mental properties. More on this in 1.3 below.

P5. $\forall u \forall \mathrm{M} \in M\{\mathrm{M}(u) \rightarrow \exists w \exists x[\mathrm{W}(w).\mathrm{P}^{\mathrm{w}^-}(w,\mathrm{w_a}).\mathrm{C}(x,u).\mathrm{I}(x,w).$
$\forall y \forall z\{\mathrm{W}(z).\mathrm{I}(y,z).\mathrm{C}(y,u).\mathrm{P}^{\mathrm{w}^-}(z,\mathrm{w_a}) \rightarrow \mathrm{M}(y)\}]\}$

This tells us that any individual *u* in the actual world $\mathrm{w_a}$ with a mental property M has a counterpart *x* at a minimal physical duplicate *w* of $\mathrm{w_a}$, and any counterpart *y* of *u* in any minimal physical duplicate *z* of $\mathrm{w_a}$ also has M. Now while the existential quantification was necessary in P4, it is not so here. This is because we know that actual individuals are physical, and so will have counterparts at worlds that minimally physically duplicate $\mathrm{w_a}$. In P4, you may recall, we needed the existential quantifiers to prevent the definition being vacuously true of worlds that do not contain anything physical. If we are only defining physicalism for the *actual* world, no such caution is necessary. But then we can rewrite P5 thus:

P6. $\forall u \forall \mathrm{M} \in M\{\mathrm{M}(u) \rightarrow \forall y \forall z[\mathrm{W}(z).\mathrm{I}(y,z).\mathrm{C}(y,u).\mathrm{P}^{\mathrm{w}^-}(z,\mathrm{w_a}) \rightarrow \mathrm{M}(y)]\}$

P6 tells us that any actual individual with a mental property M has a counterpart at any minimal physical duplicate of $\mathrm{w_a}$, which also has M. Because, like P2, P6 says nothing about worlds that are not minimal physical duplicates of $\mathrm{w_a}$, it has both the dependency and wayward atom problems. But P6 does not entail P2 as things stand, nor should it. This is because P6 defines physicalism only for mental properties, whereas P2 defines physicalism *simpliciter*. How are we to understand the relationship between the two? A natural way to think of the matter is to try to define a notion of world duplication with respect to the set *M* of properties (in this case, of course, mental properties) for which P6 defines physicalism. We could then express P6 as the thesis that any minimal physical duplicate of $\mathrm{w_a}$ is a 'mental duplicate' of $\mathrm{w_a}$. Problematically, however, there is no obvious way of understanding mental duplication for worlds in terms of individuals that are mental duplicates. Duplication, we are supposing to be analyzable in terms of shared intrinsic properties, and mental properties, as we have seen, are plausibly *ex*trinsic – individuals that are duplicates

*simpliciter* might differ mentally, if certain externalist theories of mind are true.[23] It is difficult to see how to formulate a relative notion of world duplication for mental properties; fortunately, no such account is needed. A better way to understand the relationship between P6 and P2 is to focus on the fact that P6 expresses a necessary, but clearly not sufficient, condition for duplication *simpliciter*, hence for P2. A world $w_x$ that is a minimal physical duplicate of $w_a$, as we know, contains all and only physical duplicates of $w_a$ individuals. Now if $w_x$ is to be a duplicate *simpliciter* of (or perhaps better, *identical to*) $w_a$, then those individuals will need to have all the *other* properties their $w_a$ counterparts have as well, intrinsic or extrinsic; its mental properties are, of course, a subset of these. As I said, it is no surprise that P6 doesn't entail P2, as P6 quantifies only over mental properties – in its present form, P6 defines physicalism for a subset of the properties covered by P2. But now if we allow *M* in P6 to include *all* non-physical properties, then it follows that if physicalism (now defined *simpliciter*) is true according to P6, then our physical duplicate counterparts at a minimal physical duplicate world $w_x$ of $w_a$ will have *all* the same properties as we do, intrinsic or otherwise. Which is to say that $w_x$ is a duplicate *simpliciter* of $w_a$. Hence if P6 is rewritten so as to define physicalism *simpliciter*, P6 entails P2.

The converse entailment is a much simpler matter. According to P2, if physicalism is true, then a minimal physical duplicate $w_x$ of $w_a$ is a duplicate *simpliciter*. But by definition of duplication *simpliciter*, it follows that $w_x$ contains all and only individuals that are duplicates of $w_a$ individuals, and which also have all the same relational properties as their $w_a$ counterparts. But from this it follows that any counterpart at $w_x$ of a $w_a$ individual has all the same properties, intrinsic or otherwise. Whatever the set of properties for which P6 defines physicalism, then, P6 will be entailed by P2. If P6 defines physicalism for a *subset* of properties, then P2 entails P6 *a fortiori*; if P6 defines physicalism for all properties (i.e. physicalism *simpliciter*) as does P2, then the two are equivalent. So in conclusion, P3, if not (implausibly)

[23] Putnam's [1975c] 'Twin-Earth' thought experiment, on the reasonable assumption that the content of intentional mental states is the same as the content of their propositional objects, is an argument in support of this very point. See my discussion of the problem of relational properties above for more on this.

restricted to intrinsic properties, and with the (implausible) wide-scope modality removed, is more-or-less equivalent to P2. P3 solves the dependency and wayward atom problems only at the expense of implausibility on other grounds. I will proceed to argue that those problems are not as serious as they might seem, so that global supervenience theses such as P6 and P2 are, after all, plausible definition of physicalism. In what follows, for reasons of exposition, I formulate my arguments in terms of P6. It should be clear how what I have to say applies *mutatis mutandis* to P2. The structure of the next section is as follows: first, I discuss first the wayward atom problem, then the dependency problem; I then proceed to consider the way in which P6 captures the thought that the mental is nothing over and above the physical.

### 1.3. Is global supervenience adequate?

*First*, let us reconsider the wayward hydrogen atom problem. This objection, you will recall, is that global supervenience does not account for the determination of the mental by the physical. It seems intuitively clear that hydrogen atoms *around here* can do whatever they like without affecting the actual distribution of mentality. The objector, it seems, wants physicalism to *entail* this fact. We can put the point like this: physicalism should entail the falsity of counterfactuals such as 'if this hydrogen atom were in a slightly different place, then nobody would have any mental properties'. But our definition P6 is consistent with the truth of such counterfactuals, for P6 allows that the closest possible worlds in which the hydrogen atom is displaced are worlds with no mentality at all. Paull and Sider [1992], however, present the objector with the following dilemma: either (i) mental properties are intrinsic, in which case the world described is not consistent with global supervenience, despite *prima facie* appearances to the contrary; or (ii) mental properties are extrinsic, in which case the world described *is* consistent with global supervenience, but this is as it should be. Motivating horn (i) requires an alternative formulation of global supervenience, viz. 'A *globally supervenes* on B iff any two worlds with the same distribution of B-

properties have the same distribution of A-properties as well'.[24] Appealing to this form to define physicalism *simpliciter* (rather than physicalism about a restricted domain) gives us a universally quantified version of P1:

> P1$^{\forall}$. Physicalism is true iff any two worlds that are physical duplicates, are duplicates *simpliciter*.[25]

As with P1, such a thesis is too strong, due to the problem of extras. P1$^{\forall}$ entails *a fortiori* that no world that is a physical duplicate of the actual world contains anything *non*-physical. I will not comment further on the suitability of this form of supervenience for defining physicalism, as horn (i) is implausible anyway due to the requirement that mental properties be intrinsic; still, for the sake of completeness, let us see how the argument goes. The wayward atom objection claims that P1$^{\forall}$ is consistent with the existence of a world $w_h$ that is a duplicate of the actual world save that a single hydrogen atom is displaced, and at which there is no mentality. Now consider $w_a^-$ and $w_h^-$, the actual and displaced atom worlds respectively, but with the troublesome atom removed. On the assumption that mental properties are intrinsic properties of things with minds, then according the present way of defining intrinsic, it follows that no individual at $w_a$ differs mentally from its counterpart at $w_a^-$; and no individual at $w_h$ differs mentally from its counterpart at $w_h^-$. The reason is simple: the intrinsic properties of those individuals are those properties they have regardless of what else exists; clearly the removal of a lone hydrogen atom from each world does not affect the intrinsic properties of any individuals there. From this it follows that no individual at $w_h$ has any mental properties, while those at $w_a^-$ have their $w_a$ mental properties. But $w_a^-$ and $w_h^-$ are by definition physical duplicates. Therefore we have a violation of P1$^{\forall}$, which means that the latter supervenience thesis, on the assumption that mental properties are intrinsic, is not consistent with the possibility of $w_h$. While I do not agree with the letter of this argument, I do think that Paull and Sider have an

[24] Paull and Sider [1992] p.834. Their argument, of which I present a somewhat simplified version, occurs at pp.841-6.
[25] I do not attribute this supervenience thesis to Paull and Sider, and include it here for expository purposes only. I think it likely that they would wish to restrict to domain of A to mental properties.

important general point. *Prima facie*, $w_h$ does not falsify global supervenience. But as Paull and Sider point out, universally quantified theses such as $P1^{\forall}$ are true, if they are true, for entire domains of worlds. From the putative possibility of $w_h$, on the assumption that mental properties are intrinsic, we can deduce that $w_a^-$ and $w_h^-$ are possible, and together this pair are direct counterexamples to the supervenience claim. Now to the second horn, which can be motivated without appeal to the rather dubious $P1^{\forall}$.

If mental properties are extrinsic, then we can no longer derive counterexamples to global supervenience from the putative possibility of $w_h$. This is because it was the intrinsicality of mental properties that justified the claim that $w_h^-$ and $w_a^-$ do not differ mentally from $w_h$ and $w_a$ respectively. The key point to note about extrinsic properties in the present context is that by definition, their instantiation by an individual depends on the way things are *outside* that individual. We have already seen examples of extant theories of content that have exactly this consequence – for instance, Burge's social externalism makes mentality dependent on the existence of linguistic conventions; Putnam's natural kind externalism makes natural kind thoughts dependent on the intrinsic natures of those kinds. But now as Paull and Sider point out, the fact that P6 is consistent with the possibility of $w_h$ just means that P6 does not rule out a theory of mind according to which mentality depends on the precise location of a particular hydrogen atom. Suppose that counterfactual 'if this hydrogen atom were in a slightly different place, then nobody would have any mental properties' is true; its truth does not refute *physicalism*, it merely entails that much of what we believe about the connection between mind and body is mistaken. Physicalism as defined by P6 is consistent with the dependency of mentality on *any* physical properties; and this, I maintain, is a virtue, not a vice. Physicalism, after all, is not supposed to be a *theory* of mind.[26] This is the central reason why we could not

[26] See Stalnaker's [1996] pp.229-30 for a very similar response to the wayward hydrogen atom problem. For instance, Stalnaker agrees "…that no sensible materialist would accept the possibility…[of $w_h$]. But sensible materialists are not only materialists, they are also sensible; one should not define materialism so that there cannot be silly versions of it." Nicely put. It is worth noting that the present state of science in fact does suggest that worlds such as $w_h$ are physically *impossible*. The laws of physics just don't seem to permit lone atoms to exert such a powerful influence on the

to restrict *P* in P3 to intrinsic properties – with that restriction, the hydrogen atom problem goes away, but we then place undue *a priori* constraints on empirical and philosophical theorising about the mind. If our liberalism about *P* is well-motivated by the desire to remain neutral *qua* physicalist as to particular theories of mind, then the subsequent consistency of our definition of physicalism with the possibility that tiny physical differences might make large-scale mental difference to a world ought not to trouble us. If mental properties did depend on the position of lone hydrogen atoms, then mental properties would *ispo facto* depend on physical properties, albeit in a rather odd way. Correspondingly, mental properties would still be determined by physical properties, but by properties we presently (not without good reason) take to have no bearing whatever on the mind. The *prima facie* implausibility of wayward hydrogen atom worlds is not a consequence of physicalism, nor should it be; rather, it is a consequence of the empirically well-supported (but possibly mistaken) view that brains *are* relevant to the determination of mentality in a way that lone hydrogen atoms are *not*. Precisely what properties are in the supervenience base for mental properties is a matter for theory of mind; physicalism merely informs us that those properties are some subset of the available physical ones. This concludes our response to the wayward atom problem.

*Second*, we reconsider the dependency problem. P6 is consistent with the existence of very close neighbouring worlds at which physicalism is false. As I argued above, while P4 rules out such worlds (this, we saw, was how P4 differed from P2), P6 does not. The reason for removing the wide-scope modality from our definitions is that ghosts aren't impossible; but now P6 places them closer to actuality than we might wish. Before addressing this problem directly, I will consider a similar problem raised by Witmer, who worries that P6 makes physicalism 'lucky' if it is true.[27] In essence, his objection is that physicalism could be true at a world just because the ghosts *de*

---

global distribution of mentality, or anything else, at a world. My point here, like Stalnaker's, is just that physicalists need not concern themselves with such matters. There is no *reason* why a definition of physicalism should encode substantive theses about the way the world works, however silly the denial of these theses may seem.

[27] See Witmer [2001] pp.65-9.

*facto* never get around to showing up. Witmer argues that P6 is consistent with the truth of counterfactuals such as: 'If Desmond had remembered to shave yesterday, a ghost would have appeared in his mirror to congratulate him'. The problem is that we don't want physicalism to be true at a world just because, *de facto*, the antecedent conditions of such counterfactuals are not met. Since, as Witmer points out, these antecedent conditions are propositions whose truth would not falsify physicalism, and are true at very close possible worlds, we can't allow that their truth alone would be sufficient for the truth of 'physicalistically unacceptable' propositions. If physicalism is true at the actual world, then it can't be a matter of luck that it's true – it seems that what we need, then, is a definition that entails that such counterfactuals are *false*. For the counterfactual 'if Desmond has remembered to shave, a ghost would have appeared to congratulate him' to be true, it must be the case that worlds at which Desmond remembers to shave and a ghost appears, are not further from actuality than worlds at which no ghost appears. That is, if it takes a larger departure from actuality to make the antecedent true and the consequent true than it does to make the antecedent true and the consequent false, then the counterfactual is false. As usual, let $w_a$ be the actual world, let $w_s^-$ be the closest world to actuality at which Desmond remembers to shave and *no* ghost appears, and $w_s^+$ be the closest world to actuality at which Desmond remembers to shave and a ghost *does* appear. Why does Witmer suppose that P6 tells us nothing about which of $w_s^-$ and $w_s^+$ is closest to $w_a$?

Witmer's point seems to be that P6 is consistent with there being a peculiar law of nature L that holds at $w_a$ whose antecedent condition is *de facto* never met. (We may suppose for the sake of argument that Desmond's remembering to shave is within the scope of L's antecedent.) Now if L is true at $w_a$, then $w_s^+$ will be closer to actuality than $w_s^-$, for L is violated at the latter but not at the former. Hence if L is true at $w_a$, then the troublesome counterfactual is true as well, and the brand of physicalism defined by P6 will be lucky. We can make the same point in a less bizarre way: perhaps there are physical conditions that would, if they obtained, lead to the evolution of Cartesian spirits. The trouble is that we do not want physicalism to be true at a world whose total history is such that such conditions *de facto* never happen.

Fortunately, pace Witmer, P6 rules that physicalism is false at such worlds – for as we have defined minimal physical duplication and duplication *simpliciter* – in particular condition (iv) in each definition, relating to laws – P6 entails that if L is true at $w_a$, then since L is not true at all physically possible worlds, physicalism here is false. The reason is simple: according to our definition of minimal physical duplication, L will not hold at any minimal physical duplicate *w* of $w_a$. But then according to our definition or duplication *simpliciter*, *w* will not be a duplicate *simpliciter* of $w_a$ as *w* has a differing set of natural laws. Similarly, of course, if physicalism *is* true at $w_a$ then L is *not* true at $w_a$, so that $w_s^-$ will be closer to actuality than $w_s^+$ (the latter must contain an extra law or a miracle that makes the ghost appear) making our problem counterfactual false. Thus I maintain that P6 gets the Desmond counterfactuals exactly right; why then does Witmer not see it? Here is why:

> Jackson explicitly includes the physical laws in [the] recipe [for making minimal physical duplicates] but I wish to exclude them, because I want to keep it clear that the worlds over which we are generalizing are physically possible worlds. This would be implied by the meaning of a minimal physical duplicate if we kept the laws in the recipe, but it would not be as salient.[28]

By the same token, natural laws do not appear in Witmer's conception of duplication *simpliciter* as indistinguishability in all respects. As a result, Witmer has to do a lot of manoeuvring in order to make $w_s^-$ closer to $w_a$ than $w_s^+$. In particular, he seems to want to argue that *all* the laws that hold at the actual world are true at all physically possible worlds where certain physical conditions obtain. It is far from clear to me that this is true, and equally far from clear to me why, if it is true, it solves the problem of luckiness. No matter, for P6 defined as I have defined it solves the problem without the need for any wriggling.[29] Some actual laws of nature may not be physically necessary; so much the worse for physicalism if this is the case.

[28] Witmer [2001] p.65. Witmer's preference here is stylistic, but if (as I maintain) there are *theoretical* advantages to thinking of the recipe as Jackson and I do, then we surely must do so.
[29] I will not detail Witmer's solution here, as I find it somewhat convoluted as well as unnecessary. Those interested, and those who find my solution problematic, may consult Witmer [2001] pp.67-9.

We have thus far seen how P6 entails that physicalism is false at worlds where certain problematic counterfactuals are true; and conversely, how it entails that if physicalism is true at a world, then certain problematic counterfactuals are not. However, the more general dependency problem is unsolved: P6 does *not* rule out a world $w_g$ that is a physical duplicate of $w_a$, and would be a *minimal* physical duplicate, but for the presence of a happy ghost. This possibility does not depend on the existence of any extra laws at either $w_a$ or $w_g$. However, the manner of our response to Witmer's problem suggests a similar response to the more general problem – deny that the imagined possibility is close enough to actuality to pose a problem. The remarks that follow are intended as a suggestion of what a solution to the dependency problem would look like, not a well-worked out solution. *Prima facie*, $w_g$ is very close to $w_a$. On reflection, however, this is not true – given certain assumptions, $w_g$ is a very distant world indeed. First, consider $w_g^*$, which would be a physical duplicate of $w_a$ but for the fact that the ghost there moves Desmond's razor to confuse him. But for the ghost, we may suppose, Desmond's razor would have stayed exactly where he left it. One thing we can say for certain about $w_g^*$ is that the *causal completeness of physics* is *not* true there, for by stipulation the exact position of Desmond's razor does not have a physical cause at $w_g^*$. Therefore, if completeness *is* true at the actual world, then the world at which a ghost appears is a *huge* departure from actuality, as it requires that a general empirical truth about the actual world does not hold. If these remarks are correct, we have the promise of a way of defining the neighbourhood that $w_a$ occupies if physicalism is true – all the other worlds in the neighbourhood will be worlds where the completeness of physics is true. This looks better – some physically possible worlds will be such that physics is causally complete, other not, in virtue of containing mischievous ghosts.

On its own, however, the completeness of physics will not do the work we have in mind for it. We need also to include the proposition that there widespread overdetermination is unthinkable in the condition that defines the relevant neighbourhood. This is because ghosts whose mental properties have (either actual or potential) physical effects will not violate the completeness of physics at worlds

where those effects *also* have *physical* causes. Further, ghosts whose mental properties are *epiphenomenal* will be *incapable* of violating completeness; we therefore need to add into the neighbourhood-defining proposition the premise that mental events (at least potentially) have physical effects. Now the set of propositions that defines the relevant neighbourhood is just the premise set of the causal argument! Provided the argument is a sound and valid argument for P6, we can *add* to P6 quantification over *these* worlds "for free". The resulting modality is difficult to incorporate into quantified modal logic, but much easier to define in counterpart-theoretic terms. If we so desire, we may express the dependency missing from P6 by reintroducing the wide-scope quantification we had in P4, thus: let φ = the conjunction of the completeness of physics, the denial of overdetermination, and the efficacy of the mental; let $W^{\alpha}(x)$ = '*x* is a physically possible world where φ'. Then we have:

$$P6^{\forall}. \quad \forall x \forall u \forall M \in M\{W^{\alpha}(x).M(u) \rightarrow \forall y \forall z[W(z).I(y,z).C(y,u).P^{w\text{-}}(z,x) \rightarrow M(y)]\}$$

Now there is a clear sense in which mental properties *depend* on physical properties, for if physicalism is true at $w_a$ then it is false that there are close possible worlds containing ghosts with mental properties. The closest ghost-worlds to actuality will be worlds at which φ is false. Our ghost world $w_g$ must be a world at which either (i) there are physical events with no physical causes; (ii) widespread overdetermination is possible; or (iii) some mental properties are epiphenomenal. The truth of any of these propositions, provided they are false at $w_a$, means that $w_g$ is distant, not close. Mentality at the φ-worlds depends on the physical, for there is nothing non-physical with mental properties at any of those worlds. In the chapters to follow, for the sake of simplicity, I will consider physicalism to be defined by P6 rather than $P6^{\forall}$.

Before closing this section, I will make a few general remarks about how I take supervenience formulations of physicalism to capture the thought that the mental is nothing over and above the physical. It is unclear to me exactly how to give necessary and sufficient conditions for A's being 'nothing over and above' B. Intuition is clear

about certain cases in which A is *not* nothing over and above B; necessary conditions are easier to come by than sufficient conditions in this context. For this reason, it is much easier to make problems for a relation that purports to capture nothing-over-and-aboveness than it is to solve them. For my part, I hold that if you can duplicate the actual world *simpliciter* just by minimally physically duplicating it, then there just *has to be* a sense in which everything in it is 'nothing over and above' the physical. If P6 is true, then when God made the actual world, all He had to do was fix all the relevant physical particulars, properties and laws – and everything else took case of itself; maybe that's how come He took Sunday off. I note in passing that P3, too, captures this sense of 'nothing over and above', because it entails P6. If it is metaphysically necessary that if anything is P, then it is M, then there must be a sense in which M is nothing over and above P. There are problems, of course; in the remainder of this section, I will highlight two of them. I will respond to the first problem here; my response to the second, for expository reasons, will be postponed to 7.3.

*First*, note that as Hendel [2001] maintains, P6 entails that there is a *clear* sense in which mental properties are *not* nothing over and above the physical. The argument is simple: P6 entails that wholly non-physical beings can have mental properties; *ergo*, mental properties are not nothing over and above the physical.[30] As we have seen, however, this need not pose a problem for the dependency of mental properties on the physical, for there is a neighbourhood of worlds within which there is no having a mind without having physical properties. Might we extend this thought to respond to Hendel's objection? I think we might. Suppose for the sake of argument that role functionalism is true, and mental property M = the property of having a property that plays causal role R. Now reflect on the fact that in the φ-worlds, the only properties available as role-fillers are physical. Does it not follow that *in the φ-worlds*, mental properties are nothing over and above physical Ps that play the relevant Rs? If

[30] As we have seen, P3 does not have this problem, due to the wide-scope modal operator. The problem Hendel poses is a problem for global supervenience only – or more generally, for any supervenience thesis that allows the possibility of *non*-physical minds.

anything is ever 'nothing over and above' anything else (short of being identical to it) then *particular instances* of functional properties are nothing over and above particular instances of their realizers. But if M has *non*-physical realizers at worlds outside the neighbourhood defined by φ, then M *itself* fails to be nothing over and above the physical. I think it makes sense to say this: that 'M is nothing over and above the physical' is true for the very same neighbourhood of worlds within which M depends on the physical – true in the sense that for any of the φ-worlds, every *M-instance* is realized by a P-instance. The same applies, *mutatis mutandis*, for the metaphysical necessitation in P3 – if it is true at a set of worlds that every M-instance is such that there is a P-instance such that it is metaphysically necessary that if anything is P then it is M, then the M-instances at those worlds are nothing over and above the P-instances. Provided we can make sense of particular instances of a property being nothing over and above the physical, while the property itself *is* something over and above the physical, we can have our cake and eat it: a physicalist supervenience thesis that captures the thought that mentality is nothing over and above the physical, together with the possibility of Cartesian minds.

*Second*, there is an argument, due to Wilson, to the effect that neither P3 nor P6 captures the thought that the mind is nothing over and above the physical.[31] The argument depends on Shoemaker's *necessitarian* view of the relationship between properties and causal laws.[32] It is relatively uncontroversial that properties contribute causal powers to the particulars that instantiate them. According to one version of the Shoemaker view, properties are *individuated* by the causal powers they bestow. Assuming that causal laws describe the causal powers that properties bestow, it follows that a property is individuated by the causal laws in which it figures. Next, Wilson appeals to the possibility of emergence to show that there are properties whose existence is inconsistent with physicalism, which nonetheless supervene with metaphysical necessity on physical properties. I will describe emergentism in much more detail in chapter 6; for now, I can make do with the following: emergent

[31] In Wilson [2005], pp.433-9.
[32] See for instance Shoemaker [1980].

properties (i) supervene on the physical, (ii) are something over and above the physical, connected by synchronic bridge laws $L_E$ that are independent of the laws of physics, and (iii) have novel causal powers not reducible to the powers of their physical emergence base properties. Let P be the emergence base for some emergent E. Wilson's point seems to be that given the novel powers of E, there will be laws of nature featuring P that *depend on* the instantiation of E. That is, at worlds where $L_E$ are not true, the set of *causal* laws of nature featuring P is different, due to the fact that the powers conferred by E at $L_E$-worlds are absent. But now given necessitarianism, and on the further assumption that a property is individuated by the totality of laws in which it features, it follows that the very *nature* of P depends on E.[33] P can not be instantiated at worlds where it is not an emergence base for E. But that means that P *metaphysically* necessitates E, rather than merely *nomologically* necessitating E as many – myself included – would suppose. Further, for obvious reasons, it means that minimal physical duplicates of worlds where $L_E$ hold will have to be $L_E$-worlds as well – otherwise they will fail to be P-worlds, hence not even physical duplicates. So neither metaphysically necessary supervenience nor minimal physical duplication are adequate to capture the 'nothing-over-and-aboveness' of the mental.

For my part, I am convinced that there is something very badly wrong with this argument. One response is just to deny necessitarianism. Properties have natures that go beyond the causal powers of their instances; the powers conferred by a property at a given world are determined by its nature *together with the causal laws* that hold at that world. I will not take this route, because I wish to remain neutral at present as to the truth or falsity of necessitarianism. Another response to the argument is to deny, even given the necessitarian view, that properties are individuated by the *totality* of laws in which they figure. P contributes a certain individuative set $S_P$ of causal powers; E emerges from P and contributes an extra set $S_E$ of powers; P is individuated by the laws that describe $S_P$, and E by the laws that describe $S_E$. However, there is a

[33] The situation, I realise, is somewhat misdescribed. If the nature of P depends on E, then I ought not to refer to P's instantiation at worlds where $L_E$ do not hold. I take it nothing of import turns on this.

counter to this line of response, endorsed by Kim and arguably by Lowe, too.[34] The counter is simple: given that P contributes E, why does it not *also* contribute the powers contributed by *E*? Kim uses this line of argument in his 'supervenience argument' in support of the view that it is inconsistent to suppose that emergent properties both supervene on the physical *and* have novel causal powers.[35] If the powers contributed by E are contributed by P as its emergence base, then in what sense does E have novel powers? If P contributes the powers of E (by virtue of contributing E itself), then given the necessitarian view, the instantiation of P metaphysically necessitates the instantiation of E.[36] But E is *non*-physical, so neither P3 nor P6 defines *physicalism*, let alone captures the thought that the mental is nothing over and above the physical. This argument, although to my mind quite plainly wrong-headed, is a tricky one to refute. For reasons of exposition, I must postpone my reply until 7.3, where I will argue that emergence base properties do *not* contribute the causal powers of their emergents.

Even assuming I am right that physical emergence base properties do not contribute the causal powers of emergent properties, complications remain. Necessitarians of all flavours insist that since properties are individuated by their causal roles, the laws of nature are metaphysically necessary. Whatever set of causal powers a given property does contribute, it could not contribute a distinct set of powers while remaining the same property. It follows from this that effects supervene on their causes with metaphysical necessity – but it stretches credibility to maintain that effects are nothing over and above their causes. It is not immediately obvious to me, however, that given

[34] See for instance Kim [1999a]; Lowe [2000]. I return to Lowe's views in 3.2, suggesting that he is in fact committed to something like the position I am about to describe. Lowe himself would deny that he is so committed, and in fact explicitly denies that emergence bases contribute the powers of emergents. I am not sure, however, that this position is consistent with Lowe's view that the novel causal powers of emergent propertis is not a violation of the causal completeness of physics.

[35] There are actually *several* supervenience arguments, which Kim sometimes runs together. All are directed against the efficacy of supervenient properties, but sometimes Kim depends on the completeness of physics, and sometimes not. I discuss two versions based on completeness in 5.1, and a version that does not depend on completeness in 7.3.

[36] I should point out that I do not hereby intend to attribute the necessitarian view to Kim. Rather, my point is that if we combine Kim's views about the relationship between the powers of emergents and their emergence bases, and combine it with Wilson's Shoemakerian position, then we can conclude with Wilson that physical properties sometimes metaphysically necessitate *emergent* properties.

necessitatianism, certain *prima facie* metaphysically distinct particulars are not in fact *in*distinct – effects, we might maintain, *really are* nothing over and above their effects. After all, 'B is nothing over and above A' is supposed to express the thought that given A, we get B "for free". But if having a certain effect E under certain circumstances is part of the existence condition of a cause C, then *given* that C occurs under those circumtances, surely there's a sense in which we *do* get E for free? I doubt many necessitarians would be prepared to bite this particular bullet, but that doesn't mean it isn't a consequence of their position. I have no fixed opinion on this matter at the time of writing. A more natural response to this problem is to insist that supervenience conditionals that entail that the supervenient property is nothing over and above its base must be not only metaphysically necessary, but also *synchronic*. Effects fail to be nothing over and above their causes, for causation is diachronic. Things are not so simple, however: some philosophers maintain that *simultaneous causation* is a possibility.[37] For reasons of exposition, I return to this particular complication at the end of 1.4.

### 1.4. Sufficiency, events and properties

For the purposes of this work, I will conceive of events as 'Kim-events'.[38] On this conception, as is well known, events are not importantly different from facts, conceived as immanent particular states of affairs. Specifically, a Kim event is an object – or, more generally, a substance – possessing a property at a time. This property is referred to as the constitutive property of the event. For instance, the constitutive property of a mental event such as Bob believing that paperwork is boring at a given time, is the property of believing that paperwork is boring. We can represent such events as ordered triples [*x*,P,t], but it is important not to regard the events themselves as being ordered triples. Rather, Kim events are structured particulars, for which we may give the following existence and identity criteria.

[37] See, for instance, Lowe [2003].
[38] See Kim [1976] for details.

Existence condition: [$x$,P,$t$] occurs iff $x$ has P at $t$.

Identity condition: [$x$,P,$t$] is identical to [$y$,Q,$t$*] iff $x$=$y$, P=Q and $t$=$t$*.

I will sometimes refer to such particulars as events, sometimes as 'property-instances', depending on the context of discussion. A couple of important things to note before proceeding. First, it does not make sense, on the present account, to think of events *having* their constitutive properties; an event of a given type occurs just in case its constitutive object has the requisite property at a time. Second, although we can formulate a token identity thesis for mental and physical Kim events, we can not formulate it in isolation to a corresponding *type* identity thesis, as might those who typically endorse token identity. This is because for Kim, events have just one constitutive property – there is no room for "two" events to be the same event token and yet have unrelated constitutive properties. Token identity, on the present view, entails type identity.[39] I will have more to say about this entailment in 2.3. And third, I should make clear that I do not wish to be seen as endorsing the present conception of events. The reason I employ Kim events is that they considerably "simplify the maths" surrounding the causal argument. If you run the argument in terms of, say, Davidson events, then you have to run it twice, once for events, and again for properties. The nice thing about running it with Kim events is that you only have to run it once – an argument for physicalism about mental Kim events will have physicalism about mental properties "built-in". Token (and so type) identity is clearly one way for a mental event to be physical; what other ways are there?

It is quite common in the literature to find talk of token events supervening on others. We might say, as Kim does, that event [$x$,P,$t$] supervenes on event [$x$,Q,$t$] just in case Q is among the supervenience base for P. While I think this is a natural extension of

[39] Famously, Davidson [1970] endorses token identity without type identity. For Davidson, event identity is a matter of identity of spatiotemporal location, which means a single event can have a lot of properties. The purpose of token identity within Davidson's philosophy is, of course, to reconcile the causal efficacy of mental events with their anomalism. Mental and physical event tokens are identical, so mental events are efficacious, but mental events instantiate strict deterministic laws only under their physical descriptions, so that the mental escapes the threat of physical determinism. This is not the place for detailed discussion. Suffice it to say that if this is the primary motivation for being a token identity theorist, token identity for Kim events is going to be unmotivated.

supervenience to property instances, I will instead talk of *synchronic sufficiency* relations between token events. The matter is terminological, but will greatly simplify my analysis of the problem of the validity of the causal argument in chapter 3.4. There I will argue that if the causal argument is to establish that the mental supervenes on the physical as defined in P6, it must do so by first establishing that token physical events are sufficient (in a sense to be defined) for token mental events. In chapter 6, I show that the causal argument does not establish a strong enough form of sufficiency to license the inference to P6. In the remainder of this section, I will define the kind of sufficiency the causal argument needs to establish if it is to license that inference. The following desideratum for a theory of sufficiency seems clear: sufficiency should carry *modal force*; I will think of this in the following way: if A is *α-sufficient* for B then in any α-possible world where A exists, B exists.

The synchronic sufficiency relation I am interested in is a form of non-causal determination, and this yields a necessary condition on two events that stand in a synchronic sufficiency relation: they must share their constitutive substance. Why? Events conceived as states of affairs (objects having properties at times) must be largely metaphysically independent. How *could* [*x*,P,*t*] be non-causally sufficient for [*y*,P,*t*] if *x* and *y* are wholly distinct (in the sense of having no shared parts)? If such sufficiency relations existed, I could make it so that distant objects instantaneously changed their properties simply by making adjustments to local objects. But I can't, and neither can you, so they don't. I anticipate two objections at this point.

*The first* is that there appears to be just this kind of action at a distance in quantum mechanics. As is well known, if the Copenhagen interpretation of quantum mechanics is true, then due to the phenomenon known as entanglement, we can bring about the instantaneous collapse of a wavefunction at distant points by making local measurements.[40] The details are unimportant for my purposes; I content myself with the following two thoughts. Thought (i): action at a distance is one of the central

[40] See Hardy [1998] for a detailed but mathematically not-too-heavy description of these issues.

*problems* for quantum mechanics. It has been the motivating factor behind the search for hidden local variables that provide non-spooky explanations for the relevant phenomena; it also is one of the central motivations behind the 'no-collapse' alternatives to the Copenhagen interpretation.[41] Thought (ii): Even if spooky action at a distance happens at the quantum mechanical level, it doesn't happen with ordinary macro level events. And my point is that if there were relations of synchronic sufficiency between events whose constitutive substances were distinct, then we should expect the opposite.

*The second* objection I anticipate concerns events that involve a substance having an extrinsic property, such as being married. To take Geach's example, Xanthippe becomes a widow at the instant Socrates dies. Geach terms this a 'Cambridge change', which is understood to be a change in what can be truly predicated of an individual without any corresponding change in the individual's intrinsic properties.[42] So-called 'Cambridge events', it seems, are just the sort of events you can bring about at a distance. But this makes it look, contrary to my claim, that there are events with wholly distinct constitutive substances that stand in synchronic sufficiency relations – for instance, the event 'Socrates being alive at t' and the event 'Xanthippe not being a widow at t'. I am prepared to agree that this is the case for events (if such there be) whose constitutive properties are 'mere Cambridge properties'. Clearly, however, the central case of this work – namely the relationship between mental and physical properties – does not involve mere Cambridge properties. Events whose constitutive properties are intrinsic properties of their objects do not stand in synchronic sufficiency relations if the objects are wholly distinct. Cambridge properties do not affect the causal powers of individuals in any way – you could not, for instance, build a detector to determine, just by examining Xanthippe, whether or not Socrates is alive (provided, of course, she herself does not know). So I am prepared to limit my claim of shared substance to intrinsic properties, which, as I said in 1.1, I think of as do

[41] The classic argument that quantum mechanics needs hidden variables is Einstein, Podolsky and Rosen [1935]; Bohm [1952] develops a hidden variable approach; for a defence of the no-collapse view, see Papineau [1996].
[42] See Geach [1969] pp. 42-64.

Lewis and Langton [1998]. It may be objected that mental properties are not intrinsic, as mental contents are broad. I reply that to the extent that mental contents are broad, they are inefficacious. Indeed, belief in the causal efficacy of content is the central motivation for *denying* that content is broad. I do not wish to endorse either view here. What I will say is that I am only concerned with those *parts* of mental properties that *do* cause behaviours; those parts, I maintain, *are* intrinsic, and as such, will not stand in synchronic sufficiency relations to events whose substance is wholly distinct. I will say more about the complications raised by broad contents in 3.4.

One way for two events to share their substance is for the two events to be the possession of two distinct constitutive properties by the same object at a time. Now this gives us a simple way of thinking about the synchronic sufficiency of an event for another: [*x*,P,*t*] is sufficient for [*x*,Q,*t*] just in case P is sufficient for Q. Clearly, this is exactly analogous to Kim's criterion of event supervenience. The simple view is too simple, though, for P and Q will frequently be instantiated in different individuals.[43] To see this, consider the thermodynamic property *temperature*, whose value in ideal gases is given by the statistical function: $T = k[Nm\langle c^2\rangle]$ where m is the mass of each molecule composing the gas, $\langle c^2\rangle$ the 'root mean square' velocity of the molecules (found by squaring the value for velocity of each molecule, taking the average of the squares, then the root of the average), N the number of molecules in the gas, and k an arbitrary constant I made up to simplify matters (actually the product of several other constants). A cloud of gas being at a given temperature at a time will be a Kim event – the cloud's possession of the property of being at T, say – but will also be composed of many other Kim events – namely the molecules that compose that gas having their individual velocities. How are we to understand the relationship between composing and composed event?

At this point, I introduce the term 'aggregate' to denote a mereological sum $S_M$ of events such that (i) $S_M$ has all its components essentially; and (ii) $S_M$ essentially

[43] See Gillett [2002] for an argument to the effect that dispositional properties like hardness, and the microphysical properties that realize them, are not instantiated in the same individual.

possesses the structural property $P_S$ formed by combining the constitutive properties of its components and their spatial relational properties insofar as these latter involve only other components of $S_M$ or aggregates thereof as relata. I do not claim that this usage of 'aggregate' accords with accepted philosophical usage; no matter – I will use it to denote fusions of parts that satisfy (i) and (ii). A consequence of my definition is that you can change neither the parts of an aggregate, nor their configuration with respect to one another, without forming a new aggregate. They are, if you like, *maximally fragile* mereological sums, in the sense that they can not survive any internal changes without ceasing to be. Aggregates are composed of other aggregates (provided we accept single-component aggregates as a degenerate case) and an aggregate of two aggregates is itself an aggregate. Note that I am *not ontologically committed* to such aggregates. If you object to them, then recast what I have to say in terms of plural quantification – talk instead about *those* events and *their* properties.[44] I introduce them here merely as an heuristic device to save me talking in that way, and nothing in what I have to say depends on its being the case that an aggregate of events is itself an event.

Now, consider the aggregate formed by the molecules of a gas cloud G that is at a certain temperature T. The structural property of this aggregate is clearly *sufficient* for the property of being at a given temperature T. First, there is a clear sense in which the temperature of a gas is nothing over and above the velocities of its molecules; and second, in any physically possible world where an aggregate of molecules $S_M$ has $P_S$, there will exist a gas cloud composed of those molecules, which will be at T. From this it follows that the event (or events) represented by $[S_M,P_S,t]$ is (are) sufficient for the event $[G,T,t]$. However, conceiving the cloud of gas as a *particular* means that there is pressure not to *identify* it with the aggregate. Due to its possession of $P_S$ essentially, the aggregate is more *modally fragile* than the cloud. The cloud, arguably, can survive rearrangement of its parts, but the aggregate by definition can not. If we take modal properties such as these seriously, then it seems a straightforward

[44] See for instance Boolos [1984].

application of Leibniz's law to aggregate and cloud shows that they can't be identical. The familiar way of understanding this relationship is to say that the aggregate materially constitutes (but is not identical to) the cloud, in much the same way that lumps of clay constitute (but are not identical to) statues.

There is an interesting difficulty here that deserves mention, known in the literature as the 'grounding' or 'supervenience' problem.[45] The problem is this – unless we endorse *sui generis* modal properties, then modal properties must supervene on non-modal properties. But aggregate and cloud at any given time share all their non-modal properties (to put it in Olson's terms, they are qualitatively indistinguishable), so we seem to have a violation of supervenience. An obvious response is that the modal properties of an individual supervene only on its *essential* properties, and although aggregate and cloud share all their non-modal properties, they possess different subsets of these properties essentially. It is, however, natural to define an object's essential properties as those properties such that *necessarily*, if the object exists, it has those properties, which takes us in a circle back to modality – an object's modal properties will now supervene on the apparently unanalyzable modal *fact* that there are certain properties it possesses in all possible worlds where it exists.[46]

Nothing in what I have to say demands a resolution to these issues; as I have said, treat my talk of aggregates merely as shorthand for plural talk about their components. Thought of in this way, there is clearly no pressure to regard the aggregate as identical to the cloud, for the cloud is *one*, and the composing molecules are *many*. There is

[45] As far as I am aware, the initial statement of this problem occurs in Burke [1992], but see also Rea [1997] and Olson [2001].

[46] Notice that counterpart theory offers us a nice way around the supervenience problem. Coincident "objects" are the same object – lumps are identical to statues, clouds to aggregates. The difference in modal properties is explained by the fact that any give statue has a set of lump counterparts and a set of statue counterparts, which, due to the fact that *resemblance is sortal-relative*, are not the same set. Some take this to mean that Lewis endorses contingent identity, but this is a mistake. For Lewis, all individuals are worldbound and self-identical; it follows that in all possible worlds where a given individual exists, it is identical to itself. To say that a given statue-constituting lump of clay might not have been a statue is not, on this account, to say that the actual individual that is both statue and lump might not have been self-identical – rather, it is to say that there are worlds at which the actual individual has *lump*-counterparts, but no *statue*-counterparts. See Lewis [1986a] ch.4 for extended discussion.

nothing in this approach that prevents the cloud from being modally robust compared to those events that actually compose it. For the very same cloud, we may say, could have been composed of different events. There are, of course, those who believe that the plural quantification approach can be extended to cover talk of ordinary objects like statues and lumps – terms such as these are to be construed as shorthand ways of referring to pluralities of simples.[47] I do not wish to endorse this view. I think it plausible in the case of aggregates only because there seems little independent motivation for ontological commitment to them. (Think, for instance, of an arm-movement and the complex aggregate of microphysical events that compose it. Nothing forces the view that aggregates of events are themselves events.) The same is clearly not true of ordinary things like table, chairs and statues. For ease of exposition, however, I will continue to talk of aggregates of events as if they were events.

Despite their non-identity, a gas cloud will share the substance of any aggregate that composes it at any given time. We can think of shared substance for particulars $x$ and $y$ in terms of containment relations between the sets of their parts. In particular, $x$ and $y$, the sets of whose parts are X and Y respectively, will share the same matter just in case $X \supseteq Y$ or $Y \supseteq X$, or $X-Y=\varnothing$. If $x$ and $y$ are such that $X \supseteq Y$ or $Y \supseteq X$, or $X-Y=\varnothing$ let us write $x*y$. Then we can give the following counterpart-theoretic definition of sufficiency for properties:

For any properties P and Q, P is α-sufficient for Q iff:

1. $\exists w \exists x[W^{\alpha}(w).I(x,w).P(x)]$
2. $\forall w \forall x \forall t\{[W^{\alpha}(w).I(x,w).P(x)_t] \rightarrow \exists y[x*y.Q(y)_t]\}$

This formula in (2) tells us that P-instances are synchronically α-sufficient for Q-instances just in case if any individual $x$ in any α-possible world has property P at any time $t$ then there is an individual $y$ that shares the substance of $x$ that has Q at $t$. The formula in (1) is included because without it (2) makes any property that is α-

[47] See for instance van Inwagen [1990].

impossible vacuously sufficient for any property instantiated at an α-possible world. For (1) states that it is α-possible that something is P, thereby ruling out vacuous sufficiency relations. Notice that sufficiency as defined above it is neutral between cases where the individuals that instantiate P and Q are identical and non-identical. This is because $x*y$ as I have defined it is perfectly consistent with $x=y$, for clearly in that case X-Y=∅. (For this reason, the definition is also consistent with a property's sufficiency for itself.) Further, since we can vary α according to context, we can account, *inter alia*, for the metaphysical sufficiency of the property of being red for the property of being coloured, (or the property of being water for the property of being $H_2O$); and the physical sufficiency of realizer properties for the functional properties they realize (in this case the realizer properties will need the laws of physics in order to play the role associated with the realized property). Now we may give a necessary and sufficient condition on the synchronic sufficiency of an event for another:

[*x*,P,*t*] is α-sufficient for [*y*,Q,*t*] just in case P is α-sufficient for Q.

Unlike our previous simple account, we can now understand how sufficiency relations can obtain between events where the constitutive properties are instantiated in different individuals.

The most important thing to notice now is that if we can establish that physical properties are physically sufficient for mental properties, then we will have established our supervenience thesis P6. This is because if all physically possible worlds where the actual Ps are instantiated are worlds where the actual Ms are instantiated too, then individuals at minimal physical duplicates of the actual world are going to be mental (and, of course, physical) duplicates of their actual world counterparts. But notice: we can only infer P6 from sufficiency *if the strength of sufficiency is at least physical*. Any weaker than that, and minimal physical duplication will not preserve the mental properties. Conversely, if the strength of sufficiency is physical, then nothing over and above the physical will be required in

order to make a mental duplicate of the actual world. But if we can show that physical *events* are physically sufficient for mental *events*, then we can infer the same sufficiency relation between physical and mental properties. So, I maintain that if it can be shown that physical events are physically sufficient for mental events, then we can infer P6. And the causal argument, as I will explain in chapter 3, purports to offer independent grounds for thinking that, on pain of absurdity, physical events *just have to* be sufficient for mental events.

Before proceeding, as promised at the end of 1.3, a note on synchronic causation and synchronic sufficiency. If simultaneous causation is possible, then causes are sometimes synchronically sufficient for their effects, and effects are clearly (*modulo* doubts arising from necessitarianism) 'something over and above' their causes. *Prima facie*, nothing in my definition of sufficiency rules out cases where the instantiation of P *causes* the instantiation of Q. However, I think this is a mistake. The reason I think so is that causes sometimes *fail* to cause their effects. Cause and effect are distinct existences, and other things can always get in the way of the causal process. As I have defined sufficiency, however, there is not enough slack between properties standing in a sufficiency relation for the relation to be that of causal sufficiency; if P is metaphysically sufficient for Q (let P be the property of being $H_2O$, Q be the property of being water; or let P be the property of being red, Q be the property of being coloured), then there are no P worlds that are not Q, regardless of anything else that might exist or occur. And yet – or so, at least, I am prepared to maintain – if P is cause and Q effect, then there must be such worlds.[48] So sufficiency as defined above is non-causal. The same goes, *mutatis mutandis*, for physical sufficiency. In that case, there will be no physically possible worlds where P is instantiated and Q is not, and that just doesn't look like causation to me. This is not, of course, to say that simultaneous causation does not occur. But if it does, the relationship between

[48] Notice that this argument does not depend on any particular conception of the modal status of causal laws, or the relationship between property individuation and causality. A necessitarian, for instance, could agree that there are possible worlds where token causes do not have their actual effects. Other causes might get in the way; or the circumstances under which the cause occurs might change. By contrast, nothing can get in the way of synchronic sufficiency: if P is α-sufficient for Q, then there is no α-possible world where anything is P but not Q.

simultaneous causes and effects is not the sufficiency relation outlined here. I shall return to this matter in 3.3, where I give a brief argument, based on the preceding remarks, that the sufficiency relation between physical and mental events cannot be one of simultaneous causation.

## 2. Supervenience and Reduction

The purpose of this chapter is to give an account of the relationship between reductive explanation and supervenience. The account I give leaves much to be said, but as my aims are limited, a more complete account will not be required. My aims are twofold. *First*, I want to show that there is a perfectly good sense in which reduction of a property to physical science can establish *non*-reductive physicalism about the property. To this end, I will draw on the functional model of reduction described in Kim [1998]. Kim takes functional reduction to establish type identity, which is of course in no sense non-reductive; however, it is only when combined with some of Kim's other views on causation that his preferred method of reduction has this consequence. (I list these views below, and respond to one of them; we will return to the others at various points during the present work.) In itself, however, functional reduction is ideally suited to the empirical justification of supervenience claims. This is because functional reductions entail forms of supervenience strong enough to license the view that the supervenient properties, while not identical to them, are nothing over and above the properties on which they supervene. This form of reductionism is, I take it, just what David Lewis had in mind when he said that '[a] supervenience thesis is, in a broad sense, reductionist'.[49] A reduction that fails to establish identity will not, of course, license ontological *simplification* in the sense of showing that what we previously thought as two properties are in fact one. However, a reduction that establishes a strong enough supervenience relation (for instance P6) will license the view that the reduced properties are nothing over and above those they reduce to. *Second*, I want to motivate the causal argument. It is, I claim, precisely because we lack a reductive account of mind in physical terms that we need an argument for physicalism about the mind in the first place. Since I conceive physicalism in terms of supervenience, this would be a decidedly odd claim if not supplemented with an account of the relationship between supervenience

[49] Lewis [1983] p.29. I note in passing that I do not agree with Lewis – some supervenience theses are not reductive at all. Whether or not the supervenience of A properties on B properties is reductive depends on whether or not it is strong enough to license the view that the A properties are nothing over and above the B properties. I will say a bit more about this presently.

(traditionally, of course, thought of as *non*-reductive) and reduction. Functional reductions of mental properties would, I maintain, give us very strong grounds for endorsing a physicalist supervenience thesis about the mind without the need for any additional argument. This fact has very interesting consequences for the causal argument, as we shall see in chapter 7. During the course of achieving my two stated aims, I will draw attention to some very important facts about realization, multiple realization, type identity and elimination; these facts will inform the development of later arguments. This chapter proceeds as follows: 2.1 discusses a problem for Nagel reduction, and 2.2 shows how Kim's functional reduction solves this problem. In 2.3 I explain why Kim takes functional reductions to establish eliminativism, and in 2.4 I explain why Kim is wrong about this. I conclude by giving a rough assessment of where we have got to so far with the functional reduction of psychological properties.

## 2.1. Reduction and 'bridge laws'.

The classic Nagelian model of reduction is no longer the popular choice of theory. According to Nagel, a theory T2 reduces to a theory T1 is the laws of T2 can be derived from laws of T1 with the aid of biconditional 'bridge laws'.[50] These bridge laws are needed because the predicates of the theory to be reduced will not occur in the reducing theory; bridge laws connect up the predicates and so enable the relevant deductions to go through. This, as Kim [1998] points out, is essentially a form of deductive-nomological explanation applied to theories: theory T2 can be explained in terms of T1 if T2 can be deduced from T1 with the aid of bridge laws. There are numerous problems with this model of reduction, and I will not attempt to summarise them all here. Instead, I will focus on one particular problem, which is that derivation

[50] As seen in Nagel's [1961]. There is controversy over whether Nagel requires these laws to be biconditionals – *prima facie*, it seems clear that conditionals taking T1 predicates as antecedents and T2 predicates as consequents will enable deduction of T2 laws from T1 laws just as well. See Richardson [1979] for an argument that Nagel reduction only requires conditionals that express sufficient conditions in T1 for T2 predicates; Marras [2002], however, argues that if bridge laws only give sufficient T1 conditions for T2 predicates, then T2 laws in fact cannot be deduced. Marras thinks that proper deducibility requires *replacement* of T2 predicates with T1 predicates, but does not say why he thinks this. One possible reason is that a putative T2 law derived from a specific T1 law via one-way bridge laws will have the same modal force as the T1 law, but in general (assuming T2 to be multiply realized in T1) the T2 laws will hold across all possible T1 realizations. Finally, see Kim [1998] pp.90-2 for a brief discussion of the merits of biconditionals over one-way conditionals.

of a theory via bridge laws is not sufficient for anything that deserves the name 'reduction'. My issue is not with the D-N model of explanation; in fact, I do think that Nagel 'reductions' provide explanations (of a sort) of the 'reduced' theory. Rather, the problem is that the bridge laws themselves stand in need of explanation just as much as the theory to be reduced. For instance, dualists, epiphenomenalists, emergentists and physicalists alike can all agree that there are bridge laws connecting physical and mental properties; the disagreement is over the ontological status (and, relatedly, the modal force) of these laws. If bridge laws are to yield ontological reduction, then they can not be laws that hold independently from the laws in the reducing theory. Physicalists will take the bridge laws to be true in all physically possible worlds, and hold that they are explicable in terms of basic physical laws; dualists and emergentist will take the bridge to hold in addition to, and independently of, physical law, and so will maintain that there are physically possible worlds where the bridge laws do not hold; epiphenomenalists could go either way, depending on their view of the ontological status of the mental. Proponents of any of these positions can endorse a supervenience thesis, and hold that D-N explanations of psychology can be given in physical terms. The point here is that unless the bridge laws are *physically necessary*, then minimal physical duplicates of a world at which the bridge laws hold will be worlds at which the bridge laws do *not* hold. A reduction of psychology to physical theory, then, must be one in which the bridge laws themselves can shown to hold in all physically possible worlds. The mere *fact* of a lawful correlation between mental and physical is not sufficient for any kind of ontological reduction; what is needed is an *explanation* of this fact.[51] Not just any explanation will do: what we need in order to establish physicalistically acceptable forms of supervenience is an explanation that shows why the bridge laws are physically necessary.

[51] This point is well made in Kim [1992b] pp.124-7 and [1998] pp.95-7; Beckermann [1992] p.112; and Horgan [1993] pp.577-8. Kim and Beckermann are explicitly concerned with Nagel reduction, whilst Horgan's concern is in explaining supervenience, however their central concerns are the same. All three hold, in essence, that ontological reduction demands not only on lawful correlations, but on *explanations* of these correlations as well. I will say more about these matters presently.

There is a growing consensus that only properties that can be 'functionalized' can be shown to be necessitated by physical properties and laws. Functionalization of a property involves construing it as a second order property, along the standard lines: 'P = the property of having a property that plays causal role R'. Rigidity, for instance, *just is* the property of being such as to resist change of shape; transparency *just is* the property of being such as to transmit a significant proportion of incident radiation. Now it is no coincidence that there should be a connection between functionalization and reduction. The resources available to us at the reducing level are, in broad outline, laws that tell us how things with certain properties will *behave*. If the property to be reduced can be construed as the property of having some *other* property that behaves in a certain way, then it is at least possible for us to show that the role property is physically realized. In the case of the property of being transparent to visible light, say, if we can deduce from physical theory that the material that composes a given sample does not absorb light in the visible range, then there is no *further* question whether or not the sample is transparent. We can effectively deduce whether or not certain samples will be transparent, from a functional specification of transparency, along with a physical theory of the samples in question. And crucially, we can do so without the need for any bridge laws as auxiliary premises. We can derive physically necessary one-way bridge laws relating specific microphysical structures to transparency – any conditional that takes a realizer property in the antecedent and transparency in the consequent will be physically necessary. That is, given the laws of physics, a substance with the appropriate microphysical properties can't help being transparent, as those laws determine that the microphysical properties in question play the causal role that individuates transparency.[52]

[52] The method of functionalization fits quite well with other scientifically reduced properties. For instance, thermodynamic properties such as that of being at a certain temperature can be construed (*inter alia*) as the property of being such as to cause thermometers to display certain values. A mechanical explanation of how molecular collisions affect the molecules in thermometers in the appropriate way will count as a deduction of temperature from mechanics.

Beckermann[53] and Horgan[54] both tentatively suggest that functionalizable properties are the best candidates for reductive explanation. It should be noted that Horgan's concern is not with reduction directly, but with the question what form a physicalistically acceptable supervenience thesis ought to take. Horgan claims that 'bare' supervenience theses of the kind we saw earlier are not sufficient for physicalism; rather, to confer 'materialistic respectability' on the supervenient properties, the supervenience relation itself must be "robustly explainable in a materialistically acceptable way".[55] Horgan calls this 'superdupervenience'. For my part, I hold that the 'bare' supervenience relations *are* sufficient to express physicalism, and that Horgan conflates the metaphysical question of what form a physicalist supervenience thesis ought to take with the epistemological question why anyone should believe it. Superdupervenience, for Horgan, is really just P3 above with empirical support in the form of a reduction of the mental properties; it's reduction that puts the 'duper' in 'superduper'.

Chalmers, too, holds that there is an intimate connection between functionalizability and reduction.[56] Because phenomenal concepts can't be analyzed in functional terms, he maintains, the 'hard problem' of consciousness can't be solved – the explanatory gap between physical and phenomenal concepts is here to stay. Here, in very brief outline, is how the story goes. Most philosophers agree that Jackson's so-called 'knowledge argument' shows that phenomenal concepts are not functionalizable.[57] Mary the colour scientist, as is familiar, learns all there is to know about the physical processes (including all the higher-order sciences that supervene on those processes) that realize colour perception without ever having seen anything coloured. It seems intuitively clear that when she first sees red, she learns something new – "so *this* is what it's like to see red" being the most common candidate. The disagreement

[53] Beckermann [1992] p.112-3.
[54] Horgan [1993] p.579.
[55] Horgan [1993] p.566.
[56] Chalmers [1996]. For instance, p.44: "…the possibility of this kind of [functional] analysis undergirds the possibility of reductive explanation in general." Compare Kim [1998] p.99: "Indeed the possibility of functionalization is a necessary condition of reduction."
[57] Jackson [1982].

between physicalists and non-physicalists is whether or not the something new Mary learns is a new (non-physical) fact. Both sides agree that there is a 'psychological' aspect of phenomenal redness, and that this concept can be functionalized – for instance, we might (partially) define phenomenal redness as the state normal individuals go into when they look at something red in the right conditions. By hypothesis, Mary already possesses this concept, and can deduce that she will have the phenomenal experience so defined, from the relevant physical facts – in this case facts like surface reflectance's of objects, ambient lighting conditions, and so on. However, both sides also agree that Mary gains a *new* concept when she first *experiences* phenomenal redness. If this is so, then the new concept she gains can't be a functionalizable concept.[58] Kim, for the same reasons, is also pessimistic about closing the explanatory gap between physical science and consciousness; he confines his reductionism to those properties that can be functionalized, while at the same time doubting that functional accounts of phenomenal properties can be given.[59]

## 2.2. Functional reduction

Let us agree, then, that functionalization is a necessary condition on the reduction of a property, and take a closer look at the model that Kim proposes. Functional reduction, for Kim, involves four stages. The first three are enough to establish supervenience; the fourth is independently motivated, and (mistakenly, in my view) turns realization relations into type identities. As we shall see in 2.4, the problem with Kim's argument for stage 4 is that it turns on a flawed conception of realization. Before proceeding to

[58] See for instance the deflationary response in Horgan's [1984] for agreement that Mary does, indeed, learn a new non-functional concept – one that expresses a fact she already knew in physical-functional terms. The fact expressed by her new concept (that it is like *this* to see red), while not *explicitly* physical, is nonetheless, *ontologically* physical. Similar deflationary themes are to be found in Papineau's [1998], who also sees the connection between functionalization and reduction. Papineau, however, thinks in terms of Lewis's [1966] 'argument from realization' according to which concepts associated *a priori* with a functional description are reduced by finding the physical states that play the associated role. Roughly, for Lewis 'mental state M' is by definition equivalent to 'the occupant of causal role R'. Nothing for my purposes turns on the epistemic priority of such associations.

[59] Kim [1998] pp.101-3.

discuss that issue, I will detail the first three stages, explaining how they establish supervenience.[60]

> Step 1 E must be *functionalized* – that is, E must be construed, or reconstrued, as a property defined by its causal/nomic relations to other properties, specifically properties in the reduction base **B**.
>
> Step 2 Find realizers of E in **B**. If the reduction…of a particular instance of E in a given system is wanted, find the particular realizing property P in virtue of which E is instantiated on this occasion in this system; similarly, for classes of systems belonging to the same species or structure types.
>
> Step 3 Find a theory (at the level of **B**) that explains how realizers of E perform the causal task that is constitutive of E (i.e. the causal role specified in Step 1).

The first step is relatively *a priori*, and involves the specification of the causal role that individuates the property to be reduced. The second and third steps are empirical, and jointly involve showing that E is realized in by properties in **B**. Step 1 makes it possible to *deduce* E from **B**, by establishing a conceptual link between E and causes and effects specifiable in **B**. That is, if E *just is* the property of having some property that stands in certain "causal/nomic relations" to properties in **B**, then if some property P in **B** stands in those very relations, there is no further question as to whether or not P is a realizer of E. It is important to note how important steps (2) and (3) are in this context. First, note that the mere fact that a property can be *construed* as functional does not entail that the property is physically *realized*, or indeed that it is realized *at all*. This is an important point – any property E with a typical causal role R (i.e. any *property*) will, in general, be coextensive with the second-order concept 'the property of having a property that plays R'. But it does not follow from this fact that

[60] The details of the formulation that follows are taken from Kim [1999a] pp.9-18, and Kim [1998] pp97-112, unless stated otherwise.

having E *consists in* having some other property that plays R. Cartesian mental properties, for instance, have characteristic causal roles, but are not realized by any other properties. The functional reconstrual of E required in step (1) depends only on E's having a typical causal role, and entails only that it is *possible* that E is realized in another domain of properties. Steps (2) and (3) together identify putative realizers for E, and show that they are in fact its realizers. Merely finding a putative realizer is of course insufficient; without (3) we would have only a *correlation* between a second-order functional property E and a (putative) physical realizer property P, which is no more use than a Nagelian bridge law when it comes to ontological reduction. In order to deduce the functionalized property from physical theory we need to show that it follows from **B** that P in fact plays the causal role that individuates E. The crucial point now is this: if E is individuated by its functional role, and it is deducible within **B** that P plays that functional role, then E is deducible from B without the need for bridge laws.

Now the connection between deducibility of this kind and supervenience is a simple one. Suppose for the sake of argument that **B** = physics. Let some particular instantiation of E be realized by a physical property P, and allow that there is a physical explanation of how it is that P plays the causal role that individuates E. From the laws of physics along with the instantiation of P, we can deduce that E is instantiated on this occasion – and we can do so without auxiliary premises. But this entails that the instantiation of P is *physically sufficient* (in the sense articulated in 1.4) for the instantiation of E.[61] In every physically possible world, P-instances will be sufficient for F-instances. The connection between sufficiency and physicalism, we are already familiar with; but it bears rehearsing. Any minimal physical duplicate of the actual world will (by definition) have the same physical property distribution, and

[61] Notice that I do not claim that deducibility is *equivalent* to sufficiency – rather, I claim only that deducibility of the functional property from physical properties and physical laws entails sufficiency. I remain neutral as to whether it is possible for a physical property to be physically sufficient for a functionalized property, and yet the latter fail to be deducible from the former. This will be the case if, for instance, a functionalized property is physically realized but there is no way to *show*, given the laws of physics, that the realizer property does, in fact, play the appropriate causal role. I return, briefly, to this issue in 6.1, where I discuss the metaphysical commitments of emergence.

these properties (since the laws of physics are also preserved by the duplication 'process') will have their actual causal roles. But from this it follows that any second-order functional properties realized by physical properties in actual world individuals will also be realized by their counterparts at minimal physical duplicates of the actual world. So if particular instances of functionalized mental properties are deducible within physical theory, then all counterparts of actual world individuals at minimal physical duplicates of the actual world, will have all the same mental properties as their actual world counterparts. And this is just supervenience physicalism (according to definition P6) about mental properties.

Thus, I maintain that steps (1)-(3) in the functional reduction process are ideally suited to establishing supervenience (i.e. non reductive) physicalism. These steps are clearly compatible with there being many alternative physical properties available to realize E on different occasions; thus (as, indeed, Kim maintains), multiple realization is no obstacle to functional reduction. The reduction procedure outlined above is consistent not only with different physical properties realizing E across different species, but also with different physical properties realizing E in different individuals, and even in the same individual at different times. So why not stop there, and be happy with supervenience physicalism as ontological reduction? After all, there is a clear sense in which second-order functional properties that are fully physically realized are nothing over and above the physical (again, construed broadly so as to include physical laws as well as properties), despite not being identical to any of their realizers. This option is not for Kim, however, and we shall now see why this is so.

### 2.3. Kim's eliminative reduction

The next step for Kim in the functional reduction procedure is to identify E with P. As we shall see, given that E is multiply realized (which Kim accepts) the identification of E with its realizer properties leads to elimination. It is tempting to think of eliminative reduction as a *reductio* of the theory that entails it. I will not pursue this line of argument here; rather, I will show how to undermine the argument that Kim endorses for the identification of E and P. In fact, Kim has (at least) three interrelated

argument for taking this further identificatory step. Two of them – the 'causal exclusion argument' and the 'redundancy argument' – we must postpone for later chapters.[62] The argument that concerns us in the remainder of the present chapter is based on a particular view of the nature of realization, and may be termed the 'causal inheritance argument'. Steps 1-3 above tell us that particular E-instances are realized by P-instances; the causal inheritance argument adds to this a particular and *prima facie* plausible conception of realization to show that the instances must be identified.

The argument depends on Kim's 'Causal Inheritance Principle' (CIP hereinafter), which goes like this:

> If a functional property E is instantiated on a given occasion in virtue of one of its realizers, Q, being instantiated, then the causal powers of this instance of E are identical with the causal powers of this instance of Q.[63]

Kim endorses this principle because:

> …to deny it would be to accept emergent causal powers: causal powers that magically emerge at a higher level and of which there is no accounting for in terms of lower-level properties and their causal powers and nomic connections".[64]

Set aside for the moment the question whether denial of the principle has this consequence, and grant the principle for the sake of argument. Now functional reduction (up to and including stage 3 above) of a particular E-instance to a particular P-instance (in the sense that the E-instance can be derived from the P-instance as detailed above), combined with CIP, tells us that this E-instance has exactly the same causal powers as the P-instance. Kim goes on to say that CIP "exerts powerful

[62] The causal exclusion argument is based on *prima facie* plausible premises concerning the nature of causation, in particular the claim that there is no causal *work* left for supervenient properties to do given the causal powers of their base properties, and receives a detailed treatment in chapter 5. The redundancy argument is based on a conception of what it is for a property to be novel, in particular the claim that novel properties must do causal work not done by anything else, and receives a summary treatment in 6.2.
[63] Kim [1999a] p.15.
[64] See his [1992a] p.326.

pressure to identify [the E- and P-instances]". I do not wish to take issue with the view that identity of causal powers presents such a pressure, and am willing to grant for the sake of argument that *if* the E- and P-instances have identical causal powers, *then* the instances are identical. Now given multiple realization, CIP leads, by both a direct and an indirect route, to a form of eliminativism about the mental. The direct route relies on the inference from identity of causal powers to identity of the instances themselves; the indirect route relies solely on the identity of causal powers of the two instances, and is the one Kim favours. The direct route is direct because it follows deductively from instance identity that mental designators are non-rigid; the indirect route is indirect because it does not entail elimination, but recommends it on methodological grounds. It should be noted in what follows that although I take Kim to be committed to the direct route, I do not *attribute* it to him, and can nowhere in his work find an explicit statement of it as an argument. Let us briefly consider both these routes.

*First*, the direct route. Consider: how is 'this instance of E' to be understood? In his [1999a] pp.14-5, Kim is quite clear that a 'property-instance' is to be understood as a system having a property on some occasion. That is, a property-instance is not the system that *has* the property, nor a 'trope' of that property, but *the system's having* the property at that time.[65] But this means that property instances are not metaphysically different to objects having properties at times, which is to say that 'property-instance' for Kim is just another term for 'event'. Now as we saw in 1.4, token event identity, on a Kimian conception of events entails the identity of the constitutive properties of 'those' events. The claim that the E-instance is identical to the P-instance, then, entails that E is identical to P. But multiple realization flatly denies this latter identity. If E is multiple realizable, then there must be some Q that also realizes E on some

[65] Tropes are 'abstract particulars', which some (e.g. Ehring [1999]; Robb [1997]) maintain to be the relata of causation. A trope of the property of being yellow, say, is best thought of not as something *possessing* the property of being yellow, but as something like '*this* yellowness'. The property itself is usually taken to be ontologically derivative, and understood as a resemblance class of tropes. A central motivation for being a trope theorist is that the theory promises to solve the so called 'causal exclusion problem'; the burden of chapter 5 will be to show that there is a much more straightforward way of solving that problem. Suffice it to say for the moment that whatever the merits of trope theory, it is not a theory to which Kim subscribes.

occasion, such that Q≠P. But then the identity of the E-instance with the Q-instance on that occasion entails, *mutatis mutandis*, that E=Q. And unless 'E' is a non-rigid designator, it is, of course, incoherent to maintain that E=P, E=Q, and Q≠P. In addition to this, there is the further incoherence that if E rigidly designates, then it designates a second-order functional property, which we are not free to identify with any first-order realizer property. Kim explicitly recognises this latter difficulty, and holds that mental predicates are in fact non-rigid, second-order designators of first-order physical properties.[66] Now if two individuals both possess E, but one E-instance is realized by a P-instance, and the other by a Q-instance, then the two individuals do not share a mental property at all. What they do share is the property of falling under a second-order mental concept 'E' that picks out P in one case, Q in the other. And this, clearly, is a form of eliminative reduction.

*Second*, the indirect route that Kim tends to favour. For multiple realization to be true, it must be the case that the physical realizers of E do not share all the same causal powers. Unless there are causal differences between the realizers, then the realization won't be 'multiple' at all – the central theme of multiple realization is that physically heterogeneous properties get to realize the same functional properties, and what is physical heterogeneity if not *causal* heterogeneity? But combine this with CIP, and we get the result that the causal powers of E vary according to its realization in particular instances. That is, due to their different realizations, different E-instances will clearly have different causal powers. Now take all the physically possible realizers $P_1$, $P_2$, $P_3$,…,$P_n$ of E and disjoin them. The biconditional '$E \leftrightarrow (P_1 v P_2 v P_3 v \ldots v P_n)$' is physically necessary. Heterogeneous disjunctions, Kim

[66] The solution he gives is very similar to Lewis's 'realizer functionalism', which treats mental predicates as first order definite descriptions; for Lewis E = 'the occupant of causal role R' and not, as for Kim, 'the property of having a property that plays causal role R'. If anything Lewis's strategy is the more elegant, as it is obvious that his definite descriptions are non-rigid. It is less obvious that we can make sense of predicates that appear to *rigidly* designate *second*-order properties in fact *non*-rigidly designating *first*-order properties. See Lewis [1966], [1972], [1980] for details of the realizer functionalist approach; see Kim [1998] pp.103-10 for Kim's argument for the non-rigidity of second-order functional designators.

argues, are unsuitable for framing laws, as they are not projectible predicates.[67] But given the physically necessary biconditional relating E to just such a disjunction, it follows that E isn't suitable for framing laws either. Notice that it makes no difference in the present context whether E is thought of as *identical* to the disjunction of its realizers, or merely necessarily co-extensive with the disjunction. Either way, it seems E will inherit the non-projectibility of the disjunction. In support of this view, notice that we can make the same point without appealing to disjunctions at all.[68] Put very simply, the argument may be stated like this: a given E-instance – say a $P_1$-instance that realizes E on some occasion, and from which (by CIP) E inherits its causal powers – having a causal power C does not license induction to future E-instances having C, as the next E-instance may inherit its causal powers from a $P_2$-instance, and $P_2$ by hypothesis possesses *different* causal powers to $P_1$. But a property that can't figure in laws isn't worth having. The indirect route thus recommends the elimination of E as a genuine property on the grounds that it isn't a property worth holding on to. Once again, CIP leads us to the eliminative reduction of mental properties.

There is a rejoinder to the argument that both the above routes lead to elimination, and it bears mention, as it is a position that Kim has, on various occasions, and in various forms, endorsed. Given the identification of multiply realizable functional properties with their realizers, they become "sundered into their diverse realizers in different species and structures, and in different possible worlds."[69] Why, however, can we not hold on to such properties as real but relative to the structures in which they have particular realizers? Suppose the structures in question divide up neatly along species boundaries. Rather than pain *per se*, we are left with pain-for-humans, pain-for-dogs, pain-for-Martians, and so on. Each of these properties, one might wish to claim, is a perfectly homogeneous, projectible, physical kind. Such "species-specific" type identities are endorsed by Lewis as a means to square the non-rigidity of mental

[67] Kim [1998] pp.107-9. I will not rehearse the details of Kim's argument here; we examine the problem of heterogeneity in a somewhat different context in 6.5. There, I follow Papineau [1985] in offering a teleological response to Kim's projectibility challenge.
[68] See Kim [1998] p.110 for a brief statement of this very point.
[69] Kim [1998] p.111.

designators with the evident reality of human mental properties.[70] But evidently this theory makes the reality or otherwise of human mental properties dependent on their being uniformly realized in humans. What if it turns out that your mental states and mine have *different* realizers? We are left with pain-for-me, pain-for you, pain-for-Jane, and so on – the structures relative to which mental properties are genuine and shared are now *individuals*. Or perhaps multiple realization goes even deeper than that, and Jane's pain is realized in different ways at different times. Call this kind of multiple realization 'radical' – if mental properties are radically multiply realized, then the reductions get so local as to render the appeal to structure-specificity pointless, at least insofar as it was supposed to enable us to avoid eliminativism. If it does turn out that the realization of mental properties varies between different individual, then according to CIP, mental properties will be sundered into their different realizers in those individuals. But then mental concepts will not express a property that is common to all human individuals that fall under it – there will, literally, be nothing in common to those who have a shared belied save that both fall under a second-order concept. How plausible, though, is radical multiple realization? It isn't just plausible, it's *actual*. In what follows I will argue that temperature, despite being functionally reducible and a perfectly legitimate physical property, is multiply realized as well. And as a result – so I am prepared to maintain – it too suffers elimination by CIP.[71]

Temperature, as is widely remarked, is a locally reduced property. It is, as we have seen, identical to mean molecular kinetic energy in gases, but it is identical to a different statistical function, mean maximal kinetic energy, in a solid. The difference is due to the fact that the molecules in a solid exhibit much more restricted freedom of

[70] See Lewis [1980] for details. Kim [1992a] esp. pp.322-30, endorses such a view. There, he suggests that "multiple local reductions…are the rule," and rightly argues that the suitability of this strategy for avoiding eliminativism will depend on how multiple the multiple realization of psychology turns out to be. Kim suggests that the possibility of psychological laws that quantify over humans points to the uniform realization of human psychology. The putative law 'Sharp pains administered at random intervals cause anxiety reactions', if true in humans, "is true for humans…due to the way the human brain is 'wired'…." (Kim [1992a] p.324).

[71] Since formulating the argument that follows, I have become aware of a very similar case for radical multiple realization in thermodynamics, to be found in Bickle [1998].

motion that they do in a gas. This would appear to endorse the local reductionist strategy described above. Temperature thought of as a second-order functional property capable of being instantiated in both solids and gases is sundered into temperature-in-gases and temperature-in-solids, both of which are genuine properties. This strategy won't work, however, because both temperature-in-gases and temperature-in-solids are realized by different base properties on different occasions. This is due to the simple mathematical fact that the same *overall* kinetic energy for a given ensemble of molecules can be realized by a great many different particular distributions of velocities over the molecules. Consider an ensemble of three molecules, A, B and C, each with a mass of 1 unit. The temperature of this ensemble will be $T = \frac{1}{2}\sum m\langle c\rangle^2$. Allow for the sake of argument that the result on some occasion is 18 units. It follows (I leave the reader to verify this) that the $\sum c^2 = 36$ for this ensemble. The table below shows a few ways in which the molecular velocities of A, B and C might realize this sum on this occasion.

| **A** | **B** | **C** | $\sum c^2$ |
|---|---|---|---|
| 4 | 4 | 2 | 36 |
| 3 | 3 | $3\sqrt{2}$ | 36 |
| 4 | 3 | $\sqrt{11}$ | 36 |
| 6 | 0 | 0 | 36 |

Of course, this is artificial – an aggregate of three molecules does not an ensemble make. Aggregates of such small numbers of components don't really have temperatures at all. Readers concerned by this can think of A, B and C as aggregates of a billion molecules each, and the specified molecular velocities as the average velocities of the component molecules of those aggregates. Now the crucial point is that each row in the above table represents an aggregate, whose structural property can be thought of as the molecules having the velocities specified. Suppose on this occasion that T = 18 units is realized by the aggregate described in row 1. Now once again we have two property-instances: an instance of T = 18 units in the ensemble,

and a realizing instance of the structural property. For despite being identified with mean molecular kinetic energy, temperature remains second-order with respect to the particular aggregates that realize it. The 'role' that specifies this second-order property is no longer a causal one, but is instead the condition that a mathematical function takes a specific value.[72] The property of having a temperature of 18 units is identical to the property of having mean molecular kinetic energy of 18 units. But for this ensemble, this latter property will be the property of having a property such that $\sum c^2=36$, for this gives $\frac{1}{2}\sum m\langle c\rangle^2 = 18$. And the table above shows four ways of meeting that second-order specification – four first order realizers, that is, of the second-order property of being at a specified temperature. Rest assured, there are quite a lot of alternatives.[73] Now by CIP (assuming, as before, that causal power identity for instances entails instance identity), this instance of the gas's being at temperature 18 must be identical to the particular aggregate possessing the structural property defined by row 1. But that means, *mutatis mutandis*, that temperature-in-a-gas (which we thought we had saved by local reduction) is sundered into *its* diverse realizers not only within different gas clouds, but even in the same gas cloud at different times! Reductions don't get much more local than that, nor eliminations much more eliminative.

Of course, the mere fact that temperature is radically multiply realizable does not entail that *mental* properties are. It is perfectly consistent to maintain that temperature does not survive local reduction, but mental properties do, precisely because they are *not* radically multiply realized in humans. This, I suggest, is not particularly plausible.

[72] Kim apparently does not think that anything of import turns on how the roles are specified. See for instance Kim [1998] p.115-6: "…functional properties, *as second order properties*, do not bring new causal powers into the world: they do not have causal powers that go beyond the causal powers of their first-order realizers." (My italics.) What motivates CIP is the thought that second-order property-instances are instantiated wholly in virtue of first-order property-instances that meet the relevant specification, and so whatever the powers of the former, they cannot go beyond the powers of the latter.

[73] Notice that the causal powers of these possibilities is bound to differ. They will, for instance, cause distinctive and heterogeneous microphysical changes in adjacent aggregates. Despite their causal heterogeneity, however, the aggregates in question all manage to play the causal role of temperature. I need not speculate as to how this is possible (although it is an interesting question). My argument in this section requires only that such multiple realization within thermodynamics is *actual*, and given the mathematical form of the function that defines temperature-in-a-gas, I do not see how this much can be denied.

Even if human beings in the same psychological state have the same neurophysiological properties, it is hugely unlikely that they will have the same *physical* properties. Suppose a given mental property M in humans is uniquely realized by a neural structure, N. On the basis of CIP, we must conclude that M=N. Now particular N-instances play the M role by consisting of neurones firing in a particular way, and interacting with each other in such a way as to cause the effects that define of M. The trouble here is that properties like that of being a neuronal firing are statistical in precisely the same way temperature is. For instance, let's say that neuronal firings involve the rapid diffusion of ions along ion channels. Nothing in specifications such as this one will tell us how many ions, or precisely how fast, or exactly how the velocities of the ions have to be distributed. The property of being a neuronal firing, too, will be second-order, and multiply realizable, with respect to particular aggregates of ions moving with certain velocities along ion channels. It does not matter how *similar* each of the aggregates is, for identity is an equivalence relation. If we are forced by CIP to *identify* particular N-instances with particular aggregates, then particular M-instances, by the transitivity of identity, will be identified with *non*-identical aggregates. Which is to say that it is hugely unlikely (although, I suppose, not *impossible*) that anyone will ever be in the same mental state twice.

### 2.4. Against the causal inheritance principle

My response to Kim's eliminative reductionism is that CIP is false, and that contra-Kim, its falsity does not entail "causal powers that magically emerge". I could treat CIP's eliminativist consequences as a *reductio* and simply dismiss it – as we shall see in 5.4, a very similar (and intimately related) *reductio* (the problem of causal drainage) can be run against Kim's causal exclusion argument. The trouble with this kind of line is that CIP is intuitively quite plausible, and a mere *reductio* of it will offer no diagnosis of why, despite this initial plausibility, the principle is false. To see why CIP is false, we will consider a possible objection to step (1) of the functional reduction procedure. It is not difficult to respond to the objection, but the natural response brings to light some extremely important points about the relationship

between functional properties and their realizers, points that I think Kim does not fully appreciate. The objection is simple: it is crucial to step (1) that the causal role that individuates E can be specified in terms of properties of **B**. But *how plausible is that*?

Suppose for the sake of argument that the reducing theory is physiology, and that E is a mental property – the desire to ring the doorbell with the index finger of your left hand, say. In the right circumstances, part of the causal profile of E will be that its instances cause me to ring the doorbell with my left index finger *in some way or other*. It is no part of the individuation of E that it has the power to cause me to do so in one *particular* way, rather than another. I could, for instance, carry out this action with a wide range of movements of my arm, the positions of my other fingers could vary, the force applied to the doorbell will differ from one occasion to the next, and so on. Putative physiological realizer P, however, will be a property that causes particular muscles to contract, and my body to move in a comparatively specific way. These points, of course, are not unfamiliar – it is virtually platitudinous that the causal roles that individuate functional mental properties are to be specified not in terms of particular bodily movements, but in terms of *behaviours*.

Interestingly, the same is true of temperature-in-a-gas. Now I do not wish to maintain here that the causal role of temperature-in-a-gas is different to the role of the average molecular kinetic energy, for I do accept that 'these' are the same property. However, a given instance of temperature-in-a-gas will, as we have seen, be realized by an aggregate of molecules with a particular velocity distribution. This aggregate has the power to cause very specific changes in other aggregates, through specific molecular collisions and transfers of momentum, say. But it is no part of the individuative causal role of temperature, construed as a functional property, that is has these very specific powers. Rather, it causes thermometers to display certain readings, causes pressure on containers, and so on. I think it quite plausible that in general, the causal roles of realizer properties are not the same as the roles that specify the properties they realize. The causal roles of E and P are different; how then is E to be realized in **B** at all? The

resolution of this difficulty requires an account of the manner in which realizer properties "play the causal role" associated with the functional properties they realize, one that acknowledges that the causal roles that individuate the realized properties is different to the causal role of any particular realizing property-instance.

There are many examples in the literature of just such an account; I will mention but a few. Shoemaker, for instance, holds that properties are identical to sets of "conditional powers".[74] For instance, the property of being knife-shaped is (*inter alia*) the power to cut butter conditionally on being made of wood, the power to cut wood conditionally on being made of steel, and so on. As Shoemaker points out in his [2001], this metaphysic extends in a natural way to realization: P realizes E just in case the latter is a proper subset of the former. Now Shoemaker also point out that this theory of realization can be held independently of the either the view that properties are *identical* to sets of causal powers, or are wholly individuated by the causal powers they confer. The subset theory of realization is consistent with the far less controversial view that properties *confer* sets of causal powers on their bearers. On this view, P realizes E just in case the powers conferred by E are a subset of those conferred by P. Now clearly, on this view, the causal powers of functional properties will not be the same as those of their realizers. The constitutive effect of a functional property, for Shoemaker, will be a proper part of the effect of any particular realizer. This is because particular realizers will be identical to the union of the properties they realize and some other set of powers. What such a property-instance causes will be a property-instance part of which is the constitutive effect of the realized property.

Yablo (to whose position Shoemaker likens his own) holds that mental properties are related to their physical base properties as determinate to determinable.[75] In addition,

[74] Shoemaker [1980]. Most would agree that properties *confer* causal powers on their bearers. The controversial part of Shoemaker's account is the claim that the causal powers conferred is all that individuates properties. Shoemaker actually considers two versions of this theory, one according to which properties are wholly individuated by the powers they *confer*, and a stronger version according to which properties are *identical* to the powers conferred. These matters are beyond the scope of the present work.

[75] Yablo [1992]. We will examine Yablo's theory at greater length in 4.3.

Yablo endorses the view that the causal roles of determinate properties and their determinables differ. This seems clearly right – you can build detectors for scarlet things that don't detect other shades of red; detectors for red things that don't detect yellow things; and detectors for coloured things that detect any of the above. On the proviso that the determinate-determinable relation can incorporate the realizer-role relation, then, again it follows that realizer and role properties do not have the same causal powers. Finally, there is the view of realization endorsed by Gillett. He too holds that realization is not as simple a matter as the realizer properties having the causes and effects that individuate the properties they realize, but maintains that the subset view endorsed by Shoemaker does not do justice to the differences in causal powers of realizer and realized properties. Gillett gives as an example the hardness of a diamond, and argues that the realizer properties in this case are relational properties of carbon atoms, not properties of the diamond itself. Our temperature example of 2.3 supports this view, provided it counts as a case of realization – for the realizer property there is instantiated in aggregates of molecules, and temperature instantiated in the gas. And as we saw in 1.4, there are good reasons not to identify these entities. Hardness, on the other hand, *is* a property of the diamond – carbon atoms do not cut glass, but diamonds do. Still, the properties of the carbon atoms that compose a diamond *realize* its hardness, and hardness in diamonds is functionally reducible to those properties.[76] The central point for my purposes is that hardness is individuated by the power to resist changes in shape, but realizer properties play this role by holding atoms together.

Now it seems to me that theories such as those sketched above have a common feature: broadly, they entail that realizer properties play the causal roles that individuate realized properties by causing events that are *sufficient* for their constitutive effects, in the sense detailed in 1.4. The particular realizer P of E on some occasion "plays the causal role" of E by causing an event *x* that is non-causally

[76] I should point out that Gillett would not agree with this last point, but that is because he assumes a Kimian conception of functional reduction. Gillett would, however, agree that we can explain the hardness of diamonds by reference to the laws of physics and the properties of their constituent carbon atoms. And that, for me, is a perfectly good functional reduction. See Gillett [2002], [2003] for details.

sufficient for an event *X*, where causing *X* is among the constitutive effects of E.[77] The upshot of these remarks is that in step (1) above, we should not hold that the functionalization of E requires specification of its causal role in terms of properties in **B**. This, at least in the case of mental properties, is far too implausible. Rather, we should require that the causal role of E be specified in terms of properties that *supervene* on **B** properties. And in step (3), we will, correspondingly, be looking not to explain how a specific P-instance causes a *behaviour*, but how it causes events that are *sufficient* for that behaviour. On the present view, functionally reducing mental properties is a matter of finding implementing mechanisms for causal relations between supervenient properties.[78] I am aware that these remarks leave a great deal to be said; but they ought to be sufficient for my purpose, which, as I said, is to undermine CIP. Now consider again E = the desire to ring the doorbell with the index finger of your left hand, and its realization base $P_1$, $P_2$, $P_3$,…,$P_n$. If the remarks of the preceding paragraphs are correct, then there is good reason to believe that CIP is false. The reason is simple: the causal role that individuates E is not the same as the role played by any particular $P_i$-instance. This is not in itself inconsistent with CIP, for it is the individuative role of *E itself* that differs from the roles of the P-instances, and CIP identifies only the powers of specific E-*instances* with the powers of $P_i$-instances. But let E be realized on some occasion by $P_3$. Properties confer causal powers on objects. If it is accepted that the causal role that individuates E is not the same as that of any of its realizer properties, then E will not confer the same powers on objects as $P_3$. But then how are we to avoid the conclusion that the causal powers of the E-instance *differs* from the powers of the $P_3$-instance?

---

[77] Common sense seems to suggest that, in causing *x*, P must thereby also cause *X* as well. In addition, many share the intuition that if E is to cause *X*, then the only way for it to do so (given that *x* non-causally suffices for *X*) is to cause *x*. There are good reasons, however, to doubt whether either of these intuitions is correct – we take up this matter in detail in chapter 4.

[78] There is a complication inherent in the view expressed here, which bears mention. If the causal role individuative of E is specified in terms of a property E* that supervenes on **B**-properties, then it is essential, if the functional reduction of E to **B**-properties is to go through, that the supervenience of E* is physically necessary. If, on the other hand, E* is an *emergent* property, then **B**-laws and properties *alone* will not be sufficient to explain how E's putative realizer P plays the role individuative of E. We will also need to appeal to the synchronic laws that govern the emergence of E* from the property P* that P causes. In this case, E fails to be functionally reducible to **B**, precisely because something *over and above* the properties and laws of **B** are required in order to deduce E. As we shall see in 6.3, this fact is what makes what I will term *weakly emergent* properties functionally irreducible to their emergence bases.

In fairness to Kim, he does at certain times acknowledge the causal differences I have been describing between role and realizer property-instances. Refer back to our formulation of CIP in 2.3. It is interesting to note that in an earlier formulation of CIP, Kim has (in parentheses) 'or are a subset of' after 'identical with'. Let us reformulate CIP accordingly:

> If a functional property E is instantiated on a given occasion in virtue of one of its realizers, Q, being instantiated, then the causal powers of this instance of E are identical with (or a subset of) the causal powers of this instance of Q.[79]

Let the principle so formulated be CIP'. Now CIP' is not falsified by the fact that realized and realizer properties have different causal roles. It is unclear to me whether, in fact, the subset relation is the right one to account for those differences; however, CIP' at least *promises* to incorporate them, where CIP rules them out. Still more interesting is what Kim says in fn.45, which is attached to the parenthesised part of CIP':

> Whether the principle is to be understood in terms of identity or inclusion will depend on how "realizer" is understood. On a reasonable construal if *P* is a realizer of *F*, then any stronger property *P** (say *P*&*Q*, for a nontrivial *Q* consistent with *P*) is also a realizer of *F*, and *P** may have stronger causal powers than *P*, powers that we may not wish to attribute to the instance of *F* in question. The main point, though, is that an instance of a second-order property cannot have causal powers that go beyond those of the realizing property involved.[80]

I agree with all of this. Kim and I differ in two ways. First, CIP (the principle "understood in terms of identity") cannot do justice to the difference in causal roles of realizer and realized properties; and as we have seen, such differences are common to many, if not all, role/realizer pairs. Second, CIP' (the principle "understood in terms of inclusion) does not provide motivation for identifying the instances of realized and realizer properties. Quite the opposite, in fact: if the realized property-instance lacks some of the powers of its realizer property-instance, then their *non*-identity follows

[79] Kim [1998] p.54. This is not Kim's actual wording; I have reformulated for typographic consistency. The reader may rest assured that nothing of importance was lost in the translation.
[80] Kim [1998] p.129.

straightforwardly from Leibniz's law. Further, if it turns out that functional and realizer properties are instantiated in different individuals (e.g. E in persons and $P_3$ in an aggregate of neurones), then again the E-instance will not be identical to the $P_3$-instance, since property-instances by definition can't be identified if their constitutive objects differ.

Now, it is worth taking a moment to show why elimination does not follow from CIP'. Recall our direct and indirect routes of 2.3. The direct route is easily blocked given CIP', as it depends on identification of mental and physical property-instances, and CIP' entails their non-identity. Blocking the indirect route is a little more tricky. Let the causal powers of a mental property-instance E be a subset $S_1$ of the powers of its realizer $P_1$. Let the powers of the next E-instance be a subset $S_2$ of the causal powers of its realizer $P_2$. Suppose for the sake of argument that $P_1$ and$P_2$ are the only nomologically possible realizers of E. As before, it is nomologically necessary that $E \leftrightarrow (P_1 v P_2)$. Does this not render E unprojectible, as before? It need not, but there are conditions. Specifically, if E is to be projectible, then its *instances* must all share a set of causal powers. In 2.3, we saw how CIP makes this impossible by identification of the causal powers of E- and (heterogeneous) P-instances. CIP' makes it *possible* for different E-instances to have the same causal powers, via the subset relation. However, the powers of E-instances will only be homogeneous provided $S_1=S_2=P_1 \cap P_2$. And this in turn means that on CIP', the intersection of the sets of powers of the $P_i$-instances that realize any mental property M must be (i) non-empty, and (ii) identical to the set of powers that individuates M.[81] To put the point in a less abstract way, let E=the desire to ring the doorbell in some way or other. For E to be projectible, the intersection set of the powers of all the $P_i$-instances that realize E must

[81] Clapp [2001] offers a very similar account of how it is that disjunctions can designate genuine (nondisjunctive) properties. Clapp holds that in general, a predicate F denotes a property P just in case there is some nonempty set S of powers such that $F(x) \leftrightarrow x$ has all the powers in S. Applying this to disjunctions of properties, we obtain the condition that a disjunction expresses genuine property P just in case the intersection set S of the causal powers of the disjuncts is non-empty, and anything that possesses all the powers in S is P. See also Penczek [1997] for an argument that disjunctive properties can be causally efficacious with respect to a property P only if each of the disjuncts is capable of 'standing on its own' – in other words, just in case all the disjuncts share the power to cause a P-instance.

contain (*inter alia*) the power to cause the doorbell to ring. Now suppose that $P_1$-instances have the power to cause the index finger of my left hand to press the doorbell, and $P_2$-instances have the power to cause the finger of my right hand to do so. The set $P_1 \cap P_2$ contains (*inter alia*) the power to ring the doorbell, which is as required. An attractive feature of CIP' is that the causal powers of mental property-instances are, in essence, the powers of their realizer-instances, minus the extraneous details such as which hand you ring the bell with, how hard you press it, and so on. An interesting and important question remains: *how come* the intersection set of powers of the physically heterogeneous realizer property-instances of a mental property is non-empty? (Alternatively, how come physically heterogeneous properties get to share causal powers?). In 6.5, we will examine a teleological response to this question.

## 2.5. Functionally reducing the mind

Finally, I will make a few remarks on the current state of play with regard to the ongoing functional reduction of mind. None of these remarks is particularly tendentious; however, their truth has profound implications for the causal argument, as we shall see in chapter 7. First, note that functionalization seems highly plausible for *intentional* mental properties. As I mentioned earlier, there is good reason to believe that phenomenal properties will resist construal in causal-functional terms. But properties such as beliefs are, it seems, prime candidates for reduction. On reflection, this is relatively unsurprising. Mental concepts are embedded in a folk-psychological theory that ascribes to mental properties certain characteristic roles. In addition, however, as interpreters, application of the theory is how we decide, in particular cases, whether or not people possess those properties. Just as we are reluctant to attribute to something a certain temperature if our thermometer tells us different, so we are reluctant to attribute a belief, say, to an individual for whom the available evidence suggests different. Of course, auxiliary hypotheses are available in both cases for those who really want to persist with attributing properties that don't provide causal evidence of their instantiation. But there is a clear sense in which, for the purposes of interpretation, we are all functionalists. Now I do not for one second

think that functionalism follows from our interpretive practices. Rather, my point is that there is a good fit between the two, precisely because mental properties possess characteristic causal powers. That's why interpretation based on largely functionalist criteria *works*. Once this much is admitted, it is not too great a step to the view that mental properties have their causal powers constitutively. Let's accept, then, that the first stage of the functional reduction process is largely complete. We know quite a lot about the causal powers of mental properties. As skilled interpreters, we have a head start. That isn't, of course, to say that we don't still have much to discover about psychology; but it seems to me that we know enough to get started.

We're not doing too badly with stage 2, either. Scientific study of the brain has revealed all sorts of interesting things. For instance, we know that there are correlations between certain mental activities and levels of activity in certain parts of the brain. We know that damage to specific areas of the brain is correlated with specific impairments of mental ability. We know roughly which bits of the brain are responsible for speech processing, emotional responses, dreaming, and so on. Unfortunately, at least in the simplified terms of functional reduction theory, that is where the matter ends. While we have an idea what the putative realizer properties look like, we are not even close to being able to *deduce* functionalized mental properties from the neurophysiology, biology, chemistry, or physics of brains. This is because we do not possess sufficiently complete neurophysiological, biological, chemical or physical theories of brains to explain how the putative realizers play the causal roles we associate with mental properties. Knowing that certain brain properties correlate with certain mental properties, as I have argued, is no use at all in itself when it comes to establishing that the mind is nothing over and above the physical. What we need are theories that tell us *how* the brain properties play the causal roles we specify in step 1, and this, I take it to be relatively uncontroversial, is something we just don't have right now. We can be quite confident in our knowledge of what a physical property would have to do if it were to be realize a mental property; and we can be equally confident that if mental properties are physically realized, then their realizers are brain properties of some kind. However, providing

explanations of how our putative realizers play the causal roles in question is beyond us. And this, in turn, precludes our knowing whether, in fact, the relationship between brains and minds is realization at all. As I said in 2.1, just about every metaphysic of mind thus far advanced (physicalist or otherwise) proposes a correlation of some sort between mental and physical properties. I stress these points because there seems to me to be a tendency among philosophers to suppose that the plausibility of functionalism about the mind somehow lends support to physicalism in and of itself. And as I said in 2.2, this is not so: that we can construe a property as functional tells us nothing about how the property is realized. That's one of the attractions of functionalism. Mental properties are not yet functionally reduced to anything else, and empirically, at least, metaphysics of mind is up for grabs. That's where the causal argument comes in.

# 3. Premises of the causal Argument

The causal argument depends on three crucial premises: the efficacy of the mental, the completeness of physics, and a principle of non-overdetermination. These three premises intuitively entail that the mind 'is' in some sense physical. The purpose of this chapter is to clarify the premises, and to give some justification for each, before proceeding to examine the causal argument. In 3.1, I summarise the evidence for the efficacy of the mental. I will not do much to convince anyone who isn't already convinced that the mind is efficacious that they ought to be. Rather, I simply draw attention to the fact that there is a strong body of *prima facie* evidence that mental events cause physical events. In 3.2, I focus on physicalist responses to Hempel's dilemma, and so formulate an inductive argument for the completeness of physics, based on evidence from the successes of physiology. This argument will prove of central importance in chapter 7. In 3.3, I examine what is bad about overdetermination. It is a confusion in the literature on mental causation that different authors group principles different both in content and motivation under the banner of non-overdetermination. I clarify what I take it to mean, and why physicalists should take it to mean the same as I do. In 3.4, I give a run-through of the causal argument, and show how it establishes supervenience. I show that there is room to question the deductive validity of the argument, for it leaves open varieties of supervenience consistent with non-physicalist positions.

## 3.1. Efficacy of the mental ($E_M$)

This premise simply states that mental events cause physical events. When I decide to pick up my glass, my arm moves, my hand grasps the glass. We can intervene in the physical world, change where things are located, how they move, and so on. Does it make sense to deny $E_M$? One might wish to claim that epiphenomenalism is incoherent. The relations between our mental lives and events in the world are, one might say, *paradigm case* causal relations. If the concept of 'cause' applies to anything, then it surely applies to mental events! Such arguments, however, are famously question-begging: that mental events are paradigm case causes is exactly

what epiphenomenalists will deny. However, perhaps the thought involved does contain the genesis of an argument – even if we can't argue *directly* that $E_M$ is true on conceptual grounds, there may, nonetheless, be a more circuitous conceptual route to it. It is tempting to argue, for instance, that the concept of *causation* is logically posterior to the concept of causal *explanation*.[82] If this view is correct, then a decision about whether an event or property is causally efficacious ought to be taken against a background of facts about the sort of role the entity in question plays in our explanatory practices. So, for instance, if an event figures in a singular causal explanation of some effect, then there is no *further* question about its efficacy. Explanatory relevance is what its efficacy *consists in*.

There are two lines of response to this argument. The first is to point out that it differs from the paradigm case argument only in that it begs, so to speak, a different question. Suppose it to be sufficient for the efficacy of an event that we can give a causal explanation in terms of that event. The obvious question now is, *can* we give causal explanations in terms of mental events? If epiphenomenalists are right, then presumably the mental epiphenomena will somehow (depending on your chosen metaphysic) *accompany* genuine physical causes. And the point now is, it is open to the epiphenomenalists to insist that putative causal explanations given in terms of the epiphenomenal events are merely *apparently* causal, and no more than this precisely because the events they cite are not efficacious. The second response is due to Jackson and Pettit, and directly challenges the claim of a conceptual link between causal explanation and causation.[83] The challenge takes the form of counterexamples: they claim that we can give causal explanations of an effect in terms of properties that *don't* cause it, provided those properties are (perhaps *ceteris paribus*) sufficient for the instantiation of other properties that *do* cause it. In their view, mental states are functional, their contents broad, and as such are inefficacious; such properties are

[82] See, for instance, Burge [1993]; Baker [1993].
[83] See their [1990a]. Jackson and Pettit's aim is not, I should point out, to offer such a challenge. Rather, they wish to accommodate the causal relevance of (functional) mental properties in a theoretical framework that assumes them to be causally inefficacious. However, their views are clearly relevant to the points at issue here. I consider their views further in 5.4.

nonetheless causally relevant, as their instantiation is sufficient for the instantiation of some realizer properties or other that are efficacious. This is not the place for a full discussion of this interesting theory; suffice it to say that if the theory is to some extent plausible, then to that extent it undermines the claim that the causal facts are determined by the causal-explanatory facts. For if true, it entails that all the facts about our causal-explanatory practices are consistent with the inefficacy of the properties we invoke as *explanans*.

It looks as though *a priori* arguments are out of the question; to what extent, then, is $E_M$ supported by everyday experience? Well, suppose Bob – a non-philosopher – tells you he took a break from playing the guitar because his muscles started to cramp, and he decided to rest them. If you then ask Bob whether his instantiating the property of deciding to rest his muscles was a *cause* of his taking a break, he will probably look at you funny. Still, we can surely explain to Bob the distinction between causally relevant and irrelevant properties, in the usual way: is it the colour, or the momentum, of the brick that is responsible for the window breaking? Do you think mental properties are, in relation to behaviour, more like the brick's momentum, or its colour? I think it's clear which way Bob will answer, provided he hasn't wandered off. But why? If this commitment to the efficacy of mind is *mere* intuition, then how come we all (at least those of us not persuaded otherwise by philosophical argument) share it? Consider a closely related question: why are we all so sure that the brick's impact is the cause of the window breaking? Hume famously argued for scepticism about causation understood as a 'necessary connection' between distinct events, on the grounds that we don't ever *observe* the necessity, but see merely the constant conjunction of events.[84] According to Hume, we never gain knowledge of necessary connections, but merely become conditioned, over time, to expect the effect to follow the cause. I take it we are rather less sceptical nowadays about knowledge of necessities, but Hume is surely right about the appearances. Constant conjunction,

[84] I leave open at this point what such a necessary connection consists in. My point here is that whatever causation is, it must go beyond the mere conjunctions that we observe. As is familiar, I might observe, each day of my life, the milkman turning up on my doorstep immediately after the postman, without there being a causal link between the two events.

although clearly not sufficient for causation, is just about the only evidence we have that a deeper (necessary) connection exists. Conversely, a necessary connection will often (but not, it goes without saying, *always*) count as the best explanation of *why* events of a certain type are always followed by events of another. For if there isn't a *necessary* connection between brick impacts and windows breaking, then we have our work cut out in explaining the truth of the *de facto* generalisation: 'window breakages always follow brick impacts'. If this much is granted, it seems as though the evidence for $E_M$ is on exactly the same footing as the evidence for any other causal claims we might wish to make. It's hard to see a principled difference between the evidence available in support of (i) that a brick's impact will break a window, and (ii) that a desire for water will make me go get some.

It isn't just post-hoc that we see conjunction either – folk psychology enables us to predict to a remarkable degree of accuracy just what physical states of affairs will obtain at very specific future times. There are those who deny this, but I really don't see why. If I felt like it, and had enough cash at my disposal, I could arrange to meet my good friend Owen by the gas barbecues at Coogee beach 6pm Australian time, February 7$^{th}$ 2006 for a party. He's not the most reliable person I know, but still I reckon there's a better than evens chance he'd be there. But that's a prediction, based on the interpretive attribution of mental properties alone, about the state of a very specific part of the world at a future time. Not only that, it's significantly more likely to be right than comparable predictions, made using the best available science, about what the weather is going to be like that weekend. If the psychological properties attributed in order to make predictions like these are not the properties in virtue of which Owen ends up at Coogee, then it is a very strange thing indeed that my prediction turns out to be right.

Of course, the evidence for causal connections supplied by these appearances is defeasible – sometimes we'll get epiphenomena constantly conjoined with effects they (obviously) don't cause. And in such cases, there is a perfectly good non-causal necessary connection between the epiphenomenon and the effect. Suppose, for

instance, that car engines that are about to give up tend to emit a characteristic death rattle just prior to doing so. The rattle and the end of the engine are, of course, effects of a common cause – and the (assumed) fact that they are constantly conjoined is, I take it, just one of the many reasons why constant conjunction is not sufficient for causation. The point extends to prediction, too: the engine's death-rattle significantly raises the probability of the occurrence of the events that actually do cause the engine to die – components interacting in ways that cause them to stop working, say. Then of course the occurrence of the epiphenomenon will be a good predictive indicator of the occurrence of an effect of that type.[85] But that's not a problem, as I only claim an evidential link between such conjunctions and the truth of causal claims, and no evidence *deductively entails* the truth of a proposition. What evidence does is warrant belief in a theory, and the constant conjunction of mental events and physical events is no exception. What our experience tells us is that *prima facie*, $E_M$ is true. What evidence would defeat this *prima facie* justification? In the case of our rattling engine, we can test the theory that rattles cause engines to stop – say by letting two qualitatively identical engines run, one in a vacuum, the other not, and seeing if they stop. If they both do, then we need to look for other causes of the rattling engine's failure. This is because we are apt to treat counterfactual dependency as a necessary condition for causation, and although we can't test counterfactual claims *directly*, if two engines of similar constitution behave in the same way whether the rattle is present or not, this surely suggest that the counterfactual 'if *this* engine hadn't rattled, it wouldn't have stopped' is false. Unfortunately, there is no obvious way to run similar experiments involving mental properties. What we can do, however, is look to see whether $E_M$ is consistent with other equally well supported aspects of our world

[85] Jackson and Pettit [1990a] go further, arguing that we can appeal to such epiphenomena to give causal explanations of the effects of the events whose probabilities they raise. They draw an analogy with computer programs, which, they maintain, do not cause the patterns I see now on my computer screen as this footnote, but make probable the occurrence of other events that do cause the patterns. In essence, they argue that functional mental properties are not efficacious, but that this is okay as the phenomena that really matter to us – namely the predictive and explanatory powers of mental property-attributions – can be saved without endorsing their efficacy. Functional properties get to figure in causal explanatory *laws* without being *causal*. I need not take issue with this claim here, but I will return to their views in chapter 5, where I discuss the causal exclusion problem. Denying $E_M$, as we shall see, is among the possible solutions to that problem, but it is far from being the least implausible one.

view. The key phrase here is 'equally well supported': for all I know, there may be a philosophical argument sufficient to defeat the justification for $E_M$ afforded by experience. Given the strength and availability of the evidence, however, it had better be a pretty compelling argument. In particular, if a candidate argument rests on premises or theories whose justification is *obscure*, then we should have no hesitation whatever in rejecting it as unsound. The burden of proof rests squarely on the shoulders of those who think $E_M$ is false. As I said, I have nothing to say about $E_M$ that would convince a sceptic. Rather, my purpose here has been to urge the sceptic to have reasons at least as compelling for their scepticism as the *prima facie* evidence is for $E_M$.

### 3.2. Completeness of physics ($C_P$)

This widely held view is supposed to be empirically supported, and amounts to the claim that every physical effect is sufficiently determined by prior physical causes. I prefer the following formulation:

> $C_P$: Every physical event *y* that has a sufficient cause at *t*, has a complete, sufficient *physical* cause *x* at *t*.

This version of completeness is close to what Montero terms 'Causal Closure', differing only in respect the stipulation that *x* must be a *complete* cause of y.[86] Following Papineau, I will consider a physical cause complete just in case it has its effects entirely according to physical laws.[87] We will see presently just what work this extra stipulation is doing, and why the principle is relativised to a time. Montero also outlines other completeness theses, for instance 'Strong Causal Closure', which she defines as the thesis that '[p]hysical effects have only physical causes'. The denial of this thesis is consistent with $C_P$, as the latter does not rule out the sufficiency of *non*-physical events for physical effects. However, strong causal closure appears to

[86] See Montero [2003] for details.
[87] Papineau [1993] p.22.

follow from $C_P$ and a suitable 'principle of non-overdetermination'; if every physical event has a sufficient physical cause but it is also true in general that events have at most *one* sufficient cause, then any event that causes a physical event is *ipso facto* physical. And conversely, if an event is not physical, then it does not cause any physical effects, which amounts to the strong causal closure of the physical domain. Hereinafter, when I talk of completeness, I have in mind the weaker version defined above.

A few points are in order before proceeding. *First*, in order to accommodate indeterministic causation, 'has a sufficient cause' must be read as 'has its chances determined'. $C_P$ is then the claim that to the extent that the probability of occurrence of a physical event is determined, it is determined to that extent by prior physical events. *Second*, $C_P$ is relativised to times in order to avoid an objection raised by Lowe, that $C_P$ is consistent with *x* being a sufficient cause of *y* via some nonphysical intermediary cause M.[88] Including the temporal parameter ensures that every physical event has a complete physical *causal history*. At any point in the causal aetiology of an event at which it has a sufficient cause, we can find physical events that are jointly causally sufficient for its occurrence. *Third*, and most importantly, there is the related objection, also due to Lowe, that $C_P$ is consistent with a physical event *x* that *simultaneously causes* a mental event M, and *x* and M together are causally sufficient for a physical event *y*.[89] Lowe maintains that *y* still has a sufficient physical cause, even if it true that *x* somehow requires the action of M in order to cause *y*. Appeals to times in this case will not distinguish M from *x*, since the causal relation between them is by hypothesis simultaneous.[90] For now, notice that contra-Lowe, the imagined scenario is plausibly *in*consistent with $C_P$. For *x* is in no sense a sufficient cause of *y* *according to physical laws*. The laws by which *x* causes M govern *x*'s causing *non*-physical events, and as such clearly cannot be *physical* laws. But since M is a

[88] Lowe [2000] pp.575-6.
[89] Lowe [2000] p.576-7; Lowe [2003] pp.148-9.
[90] I will argue in 3.3 that simultaneous causal relationships are problematic when applied to the mind-body relation. The reason I will give there is that the fact that there are *always* mental causes of behaviour strongly suggests a non-causal relationship between the physical causes of behaviour and the mental causes.

necessary causal condition for *y*, it follows that *x* does not cause *y* according to physical laws alone, and as such fails to be a *complete* cause. In Lowe [2000] (p.3) the formulation he terms '1B' due to Papineau has 'complete' built in as a condition on the nature of the causal relationship between *x* and *y*. However, Lowe does not consider whether simultaneous causation poses a problem for such formulations, focussing instead on formulations of $C_P$ that require only the sufficiency of the proximal cause. Despite the fact that I disagree with Lowe that the imagined possibility is inconsistent with $C_P$, I do agree that such possibilities are apt to make any special causal contribution on the part of M 'invisible' to certain kinds of empirical enquiry.[91] In particular, scientists would quite likely frame causal laws to describe situations such as this one without reference to M at all. Causal explanations such as the explanation of *y* by reference to the law '*x*s cause *y*s', Lowe says, can *appear* to be complete if the mode of enquiry is limited to observing, *inter alia*, that *y*s always follow *x*s, despite the fact that such explanations are in fact *in*complete.[92] During the course of chapters 6 and 7, we will derive this very conclusion by different means. As we shall see, what Lowe calls the 'invisibility' of mental causation has profound implications for the argument (to follow) that is supposed to justify $C_P$, and for the causal argument.

So why believe $C_P$? If $C_P$ is true, then a scientist wanting to find the cause of a physical event will never have to go outside the physical domain; the evidence for it is supposed to stem from the amount of progress scientists have made in finding explanations for various phenomena *within* the physical domain. That scientists have made such progress is undeniable; whether their progress licenses $C_P$ is another matter entirely. In this section, we will consider the problem widely known as 'Hempel's dilemma'. A good initial account of this problem is to be found in the exchange between David Papineau and Tim Crane. Papineau [1989] appeals to $C_P$ to argue that

[91] Lowe [2003] pp.150-1.
[92] I note in passing my puzzlement at Lowe's apparent endorsement of the conjunction of (i) the efficacy of M does not render $C_P$ false and (ii) a causal explanation of *y* in terms of *x* is *incomplete* due to M's role.

mental properties supervene on the physical.[93] Crane's [1991] response is that the defender of $C_P$ is faced with a dilemma.[94] Either the 'physical' in $C_P$ means part of *current* physics, or else some idealised *future* physics. If 'physics' means current physics then $C_P$ is probably false, as there are most likely causal gaps in current physical theory, requiring entities beyond those already posited to fill them. No physicist I ever met thought that current theory had the resources to explain everything. If, on the other hand, 'physics' refers to whatever future theory *is* causally complete ('PHYSICS', say), then $C_P$ is analytic. The analyticity is due to our apparent inability to describe PHYSICAL theory without reference to its completeness. After all, we don't know what entities, laws, properties, and so on, will be essential to the explanations PHYSICS provides. The obvious worry now, of course, is that if *mental* properties turn out to be necessary to fill the causal gaps in *Physics*, then the supervenience of the mental on the *PHYSICAL* will be trivial. Papineau responds to Crane by equating 'physical' with 'PHYSICAL', and arguing that the supervenience of the mental on the PHYSICAL will non-trivial provided there is good reason to think that PHYSICS will *not* include mental properties. Papineau is confident that PHYSICS will not require mental properties; during the remainder of this section, we will try to isolate the source of this confidence.

Spurrett and Papineau[95] attempt to shift the focus of debate away from the issue of interpreting 'physics' in $C_P$, and replace the completeness of physics with the completeness of the *non-mental*. This completeness thesis states that all non-mental events are sufficiently determined by other non-mental events; whatever causal holes there are in current theories about non-mental entities, they won't be filled by mental plugs. Notice that this thesis can't be used in a causal argument for *physical*ism –

[93] Papineau appeals to $C_P$ to establish that the causal powers of mental properties must depend on the powers of physical properties. He endorses the additional thesis that any difference in properties must be manifestable as (at least potentially) differential effects. His claim is that these theses together entail that there can be no difference in mental properties without a corresponding difference in physical properties. Papineau's argument will not suffice to establish physicalism, for 'no A-difference without B-difference' is a component common to *all* supervenience theses – including the ones that aren't compatible with physicalism.
[94] See also Crane and Mellor [1990].
[95] In their [1999].

rather, its use in the causal argument will tell us that the mental is non-mental. A natural term for such a position is 'non-mentalism'. At first blush, this sounds like a contradiction, but it's not, as 'non-mental' here is to be understood as non *sui generis* mental. '*Sui generis*' translates as 'of its own nature', and in the present context refers to mental entities whose mental natures exhaust their natures. A *sui generis* mental property, for instance, will be one whose instantiation does not *consist in* having any other properties; a *sui generis* mental force will be a force exerted by a *sui generis* mental event, property, substance, etc.[96] Now it seems that the causation of bodily movement is the best place to look for possible *in*completeness of the non-mental. That is to say, if there is evidence that all physiological phenomena have sufficient non-mental causes, then there is evidence that mental properties aren't essential anywhere; so if we can find evidence there of completeness with respect to the mental, we will have evidence that the non-mental is complete.[97] Papineau ends up giving just this kind of account – in his [2001] he explicitly attributes the plausibility of the completeness of physics to twentieth century advances in physiology. There, he argues that since physiology has enjoyed significant explanatory success without mental properties, there is proportionately less motivation for believing in '*sui generis* mental forces'. And if there is no need to posit such forces to explain the causation of bodily movements, then they won't be required anywhere else either.

This line of argument, although *prima facie* plausible, is rather more problematic on closer inspection. Recall the dilemma pointed out by Crane – if 'physics' in $C_P$ is defined by current theory, then we've good inductive grounds for thinking $C_P$ false; and if defined according to future theory, then how do we know what it will essentially refer to? In response, Papineau cites the absence of mental properties from

[96] Note that *sui generis* mental properties might be related to physical properties. What distinguishes them from e.g. the functionally reducible properties we discussed in chapter 2 is that by contrast, *sui generis* mental properties are something over and above the properties they are related to. More on these matters in chapter 6.

[97] Although some maintain that quantum processes are incomplete with respect to the *mental*. We lack a physical explanation of the apparent wavefunction collapse that occurs when a measurement is made. It has been suggested that *consciousness* plays a central role – it is not measurement *per se* that causes the collapse, but measurement by a conscious observer. If this were true, then quantum mechanics would be incomplete with respect to consciousness. See Wigner [1962] for an endorsement of such a position.

physiological explanations as evidence that the non-mental is complete – for if mental properties are not essentially involved in the causation of bodily movements, then where? The claim here is that evidence from physiological research supports the claim that all bodily movements are determined by preceding non-mental events or properties. However, the evidence from physiology, Gillett and Witmer[98] argue, does not mention the 'non-mental' at all – rather, such evidence could only support the completeness of the non-mental by supporting the claim that all physiological effects (including bodily movements) have physiological causes, *ceteris paribus*. The familiar clause at the end allows for the incompleteness of physiology *per se*, which is fine – we don't require the *completeness* of physiology here, but rather its completeness relative to the mental. What is problematic, though, is that 'physiological' here seems equally subject to the original dilemma as 'physical' in $C_P$! If understood in terms of current theory, then the claim is probably false – physiological theory can almost certainly be improved within its own domain. That is, although perhaps incompletable, physiology is nonetheless not as complete as it could be. And as before, if 'physiological' refers to future theory, then how do we know what properties that theory will require? Perhaps the theory will involve ineliminable reference to mental properties.

The argument Gillet and Witmer offer is not too difficult to block. It is certainly true that "…no working scientist goes through the trouble of saying that such-and-such bodily movements have 'non-mental causes'", but that does not make it *false* that those events have such causes, nor that evidence can be gleaned from physiological research to support this conclusion. All a successful induction requires is a stock of successful physiological explanations whose success depends only on appeals to entities or properties that have the *property* of being non-mental – it matters not whether physiologists refer to that property when *they* frame *their* evidence.[99] And,

[98] In their [2001].
[99] There is an interesting question here concerning the definition of 'success' in the present context. In point of empirical fact, the successes of physiology thus far are lacking in a way that turns out to undermine the justification of $C_P$. For reasons of exposition, I postpone discussion of this point until chapter 7.

given that physiological theory has not had to posit *sui generis* mental forces to explain physiological effects, it seems that the induction will be good – for to be non-mental *just is* to be specifiable in non-mental terms. We can state the crucial induction very simply, as follows: no past physiological successes have required the incorporation of *sui generis* mental properties, so we should expect that no future successes will require the mental either. Or equivalently, like this: all past successes in physiology have involved appeal to properties that lack the property of being *sui generis* mental, so we should expect all future such successes to involve properties that lack this property as well.

There is, however, a more serious problem, for similar inductions can be run to show that we should not expect *any* new property to be introduced into a scientific ontology, which is absurd. Consider a property P not yet discovered, which, if discovered, would plug some of the present causal gaps in physiology. Now by hypothesis, no previous successes in physiology have involved appealing to P – this fact, *inter alia*, is what makes current physiology incomplete. But if we are inductively justified in expecting future successes in physiology to involve the introduction of non-mental entities, should we not also expect those entities to be non-P? That is, to induce completeness of the non-mental from physiological evidence, we need the fact that all past successes in physiology have involved entities that lack *sui generis* mental properties to support the conclusion that future successes in physiology will be similarly lacking in such properties. But then doesn't the fact that past physiological successes have involved the introduction of entities that lack P support the conclusion that P will never be introduced? Since P can be any property you like, and nothing depends on physiology being the science in question, it follows that we should not expect there to be any further modifications in any sciences, and that all the ones that have happened so far have been quite surprising. Clearly something has gone wrong – far from expecting that no new entities will be introduced to theories we acknowledge to be incomplete, in fact we expect just the opposite. The problem here is that sciences that are not yet complete within their own domains – not yet as complete as they *can* be – are quite likely to require entities not

yet within their ontologies in order to be completed. In the case of physics, for instance, it is unlikely it can provide maximally complete explanations of the relevant phenomena with its present ontology. But if this is so, then we should expect future revisions to physical theory to introduce entities of a kind not introduced by any previous revisions. For this reason, I think, no induction will be possible from what past revisions in a given scientific domain have not included to what future such revisions will not include. Surely, though, you may object, we don't expect future physiological theories to introduce *sui generis* mental entities? If this expectation is to be inductively justified, then we must look elsewhere for its source. In what follows I will argue that the source of the expectation is that there are similarities between past and present theories that license inductions about the nature of future theory. I'll explain.

There are good reasons for supposing that we can negotiate Hempel's dilemma without recourse to the non-mental at all; that is, without giving up on defining PHYSICAL. The dilemma, remember, is that $C_P$ is either false or vacuous: false if defined by current physics, vacuous if defined by future physics as we have no idea what that will be like save that it will be complete. But this is a bit quick for my liking. Typical recent advances in physics, for instance, have involved the introduction of new particles, forces, and suchlike. And these must be similar in *some* ways to the *old* particles and forces and suchlike, otherwise they wouldn't be new *particles* and *forces* and *suchlike*.[100] If this is true, then the claim that we have *no idea* what PHYSICS will be like is false, and a moment's reflection on past scientific progress supports this conclusion. Arguably the biggest revolution in scientific thinking of the last, and possibly any, century, is the shift from Newtonian to Quantum mechanics. Even so, quantum particles are really quite similar, in many

[100] In his [1999], R F Hendry claims, in support of Papineau, that there is an evidential link between current theory in a particular domain and the types of entities that a future version of that theory will have to postulate. In fact Hendry only grants Papineau this only for the sake of argument. His concern is to argue that even if it is clear that mental properties will not figure in complete physics, it is far less clear that, for instance, chemical properties won't. Current physical explanations, although they do not involve explicit reference to mental properties, often do invoke chemical properties, and it's less than clear that such references are eliminable.

respects, to their superseded Newtonian counterparts. Quantum particles may not simultaneously have precise positions and momenta, but they are still located within spatial regions, and still have velocities that fall within certain ranges. And they engage in collisions, exert forces upon one another, form bonds with other particles due to these forces, and generally do most of the things the Newtonian atoms used to do – albeit in a very different *manner*. My general point here is simply that quantum particles resemble Newtonian atoms much more closely than they resemble glaciers, or strawberries, or mental contents, to name but a few. For instance, given the extraordinary predictive success of the probabilistic approach of quantum mechanics, it would be extremely surprising (some might go further – *miraculous*, perhaps) if nature were not inherently probabilistic in character. As such, it will be equally surprising is PHYSICS doesn't involve, *inter alia*, the assignment of probabilities to various outcomes. Why does any of this matter? Because now we can frame the relevant inductions without problematic appeals to what past advances in a given theoretical domain have *not* involved.

The reason Spurrett and Papineau turn to the completeness of the non-mental in the first place is to avoid making a commitment viz. the nature of PHYSICS. But that now seems too cautious – for if the preceding remarks are right, then it seems we can say quite a lot about PHYSICS after all, *without* mentioning its completeness. Now it is tempting at this stage to construct a simple induction based on the nature of current physics to the conclusion that PHYSICS won't contain mental properties. After all, *sui generis* mentality is nothing like atoms, spin, and superposition states. Why not, then, simply argue from (i) that future successes in physics will involve the introduction of entities similar to those involved in past successes, and (ii) mental properties are nothing like current physical entities, to (iii) PHYSICS won't involve appeals to mental properties? There is nothing formally wrong with this *qua* induction, but the nature of the successes referred to in (i) means that the argument contains a sampling error. The reason is that current physics doesn't even *attempt* to account for any effects that might plausibly be attributed to *sui generis* mental properties. If the successes referred to in (i) contained a large enough number of

physical explanations of the effects of mental causes, then the induction would be just fine. However, as I take to be relatively uncontroversial, the success of current quantum theory in explaining things like bodily movements is nil. Put simply, the trouble with this kind of strategy is that past successes in physics have mostly been concerned with explaining what goes on in particle accelerators and cloud chambers, and even the most diehard Cartesian would not insist that *sui generis* mental properties are needed to explain what goes on in places like *that*.[101] However, there is no reason why we can't appeal to the inductive link between present and future theory *outside of physics*. And that is exactly what Papineau does when he appeals to physiology, for physiological effects are the likeliest *explananda* to require *sui generis* mental *explanans*. The business of at least some branches of physiology is *precisely* to explain effects we know to have mental causes, and to do so in non-mental terms. The new inductive argument can be stated like this:

(i) Future successes in physiology will involve appeals to entities and properties of a broadly similar nature to those involved in past successes;

(ii) *sui generis* mental properties are *nothing at all* like anything in the ontology of current physiology; so

(iii) PHYSIOLOGY will not contain *sui generis* mental properties.[102]

Now if *sui generis* mental properties are not required to explain why our bodies move as they do, then they will not be needed to explain why *atoms* move as *they* do. If this is so, then whatever PHYSICS does contain, we can confident that it won't contain mental properties, so that the supervenience of the mental on the PHYSICAL will not be trivial. For this reason, I refer to this argument, in what follows, as the *non-triviality argument*.

[101] Melnyk [2003] runs a slightly more subtle version of the argument mentioned and rejected here. Its extra subtlety, as we shall see in 7.2, does not prevent it too from containing the very same sampling error.

[102] We will grant the physicalist this argument for the time being. Later on, I will argue that here too there is a sampling error, but this point takes some justification, and so I must postpone my argument until 7.3.

Notice that from (iii) above we can infer the completeness of the non-mental – for (iii) entails that all PHYSIOLOGICAL effects have non-mental causes. Combining this again with the thought that if we don't need *sui generis* mental properties to account for effects like bodily movements, then we don't need them for anything else either, we can infer that *all* non-mental effects have non-mental causes. Notice also that no problematic induction from what past theories have *lacked* is necessary here: provided the non-triviality argument establishes its conclusion (iii), the completeness of the non-mental follows *deductively*, on reasonable assumption. This completeness thesis, however, is no longer motivated by Hempel's dilemma, for given the non-triviality argument, the dilemma doesn't bite. If the non-triviality argument is sound, then Papineau is right to be confident that PHYSICS will not require *sui generis* mental properties. We have two completeness theses at our disposal, then – is there any reason to prefer one or the other? I think there are at least two reasons to prefer the completeness of physics.

*First*, the non-mental appears highly gerrymandered, for the only thing in common to all non-mental entities is that they lack the property of being *sui generis* mental. But if this is right, then it follows that from 'X is non-mental' we can't really infer anything *else* about X. No amount of observations of physiological entities, say, will confirm 'all non-mental entities are physiological'; the motion of tectonic plates causes earthquakes, but that fact does not justify designing an early warning system that monitors the motion of clouds. The non-mental lacks the scientific unity for the property of being non-mental to be of any real use. Don't get me wrong: as I said, I do think that the completeness of the non-mental follows from the non-triviality argument. However, it is far from clear to me what 'non-mentalism' about the mind really amounts to. As metaphysics of mind go, it certainly isn't as informative as physicalism, for all non-mentalism tells us is that the mind, whatever it is, is really something else.

*Second* the completeness of physics allows us to use the causal argument to argue for physicalism about any classes of events that have physical effects. If the argument is a good one, then it can be iterated to give a broad physicalism, not just about the mind, but about everything. Completeness of the non-X as a premise of the causal argument, however, establishes a different metaphysical claim about each X. About the mental, we will get non-mentalism; we will get non-geologism about the geological; non-sociologism about the sociological; and so on. It is unclear how, if at all, these different 'isms' should be united. All that each one says is that the X in question is really something else. The completeness of physics removes the worry – for the conclusion of causal arguments based on this premise is that everything with a physical effect is physical.

What these two arguments show is that other things being equal, the completeness of physics is a much more useful premise than the completeness of the non-mental. In chapter 7, I will argue that despite initial appearances to the contrary, current evidence does not, in fact, support the conclusion that PHYSIOLOGY does not contain mental properties, and that as such, neither the completeness of physics nor the completeness of the non-mental is supported. For the purposes of the intervening chapters, however, I assume that $C_P$ is both true and non-trivial.

### 3.3. Principle of non-overdetermination ($O_D$)

Philosophers often maintain that overdetermination is at worst anything from impossible to absurd, and at best extremely rare. I disagree. For my part, I take overdetermination to be the sort of thing that happens when two assassins shoot the same person, the bullets arrive at the same time, and each bullet would have been fatal on its own. Contrary to popular received philosophic wisdom, I think that this sort of overdetermination happens quite a lot. Your average firing squad, for instance, is set up so that more than one of the marksmen kills their victim. The reason I do not think there is anything absurd about overdetermination in the firing squad case is that the overdetermining causes are effects of a common cause (prior arrangement, the order to fire, etc.), and this is sufficient to render the co-occurrence of the multiple causes

non-coincidental. It does not matter how often overdetermination occurs, on this account – what matters is the *relationship between the overdetermining causes*.[103] Now $E_M$ and $C_P$ together entail that certain physical effects (specifically the physical effects of mental causes) always have both mental and physical causes. What really would be absurd in the case of mental causation is if there were no relationship at all between mental and physical causes sufficient to render their co-occurrence non-coincidental. For then we should have a case of widespread coincidence, which is something we really can't allow. In order to formulate $O_D$, we need to know what relationship must obtain between mental and physical causes if their co-occurrence is to be non-coincidental. In what follows, I argue that the appropriate relationship is sufficiency as defined in 1.4. I am sympathetic to much of what Kim has to say about overdetermination; I discuss his views below.

What is a coincidence? Think of the two assassins case. It has the following features: (i) both shoot the same person at the same time; (ii) the bullets strike the victim at the same time; (iii) either bullet alone would have been sufficient for the victim's death in the absence of the other. We are compelled, when faced with this sort of situation, to look for an *explanation*. The most plausible explanation is to assume some sort of collusion between the assassins; alternatively, some design on the part of a third party. If no such explanation can be supported, then we are apt to regard it as a freak occurrence. Now suppose causation were *always* like that. Either the world is full of unexplained, freak occurrences or coincidences, or else there is widespread design on the part of some agent to render the otherwise coincidental occurrences explicable. Malebranche, for instance, thought that whenever we decide to act in a certain way, God causes the right action; Leibniz, on the other hand, thought that God was too busy to be following our every move, and takes care of it all in advance. Both these

[103] Those wishing to claim that overdetermination is absurd sometimes explain away double assassin cases as cases of joint sufficiency. Events, they maintain, are *fragile* with respect to the time and manner of their occurrence, so that the victim would have *died a different death* had he been shot by just one of the assassins. This implausible theory rules out many things we would ordinarily wish to say about events, such as that they might have occurred a bit sooner, or in a slightly different way. We return to this issue in 5.5; see Lewis [1986c] for discussion.

philosophers recognised the inconceivability of widespread coincidence.[104] I take it, however, that neither position wholly satisfies.

The question remains, then, as to what relation must obtain between two causes in order for their joint causal sufficiency for an effect to be *non*-coincidental. One answer is obvious – identity. It is no coincidence that an event has the same effect as itself. It seems, however, that much weaker relations will also render simultaneous causal sufficiency non-coincidental. Consider again the two assassins case, and suppose that the person who arranged the assassination also arranged it so that the two would shoot from positions equidistant from their victim, and pull their triggers at exactly the same time. This is not freaky – as I said above, there is nothing at all wrong with the thought that assassinations are *always* like that. If you wanted to set one up, then you could do a lot worse than set it up specifically so that the poor victim's death is overdetermined. What we can't accept is the thought that assassinations might always be like that even in the absence of an explanation as to *why* they were. What is it about explanations that make certain cases of overdetermination acceptable? In this case, the actions of our assassins are effects of a common cause, and this seems sufficient to render their joint occurrence non-coincidental. For surely, given the cause – the co-ordinating actions of the person who hires them – it is not at all spooky that the assassins behave as they do.[105] What, then, of cases where it is no-one's intention that two causes overdetermine an effect? Well, in the absence of a third entity that explains their joint occurrence and efficacy, the

[104] Of course, neither Leibniz nor Malebranche were concerned with overdetermination. Rather, their concerns were with the nature of the interaction between mental substance and the world, in response to the Cartesian problematic. Both accept that $E_M$ is false, but use God to explain why it *looks true*. For Malebranche, neither mind nor body have any effects at all – bodily movements and the mental and physical states that precede (and appear to cause) them are caused by God, who can do anything He wants. For Leibniz, mental effects have only mental causes, physical effects have only physical causes, but God, who can do anything He wants, sees to it during the Creation that mental events happen immediately prior to the bodily movements they explain, and at the same time as their physical causes. Notice, however, that if global coincidences were acceptable, then the appeal to God would be unnecessary.

[105] The explanation here has exactly the same form as the Leibnizian and Malebranchian solutions to the Cartesian problematic.

two causes must, so to speak, take care of each other. Kim agrees, and we turn now to a much-quoted principle of his, the 'Causal Exclusion Principle'.[106]

Kim state this principle in a variety of (relatively similar) ways in different papers on the subject. For present purposes, the following formulation is appropriate:

> There cannot be two or more *sufficient* and *independent* causes of a single effect, except for cases of genuine overdetermination.[107]

Kim argues for this principle by showing that it has no counterexamples. The thrust of his argument is that when we consider putative cases of two causes of the same effect, then they are either insufficient, or dependent, or a case of 'genuine overdetermination'. Here, in no particular order, and slightly altered for terminological consistency, are the central cases Kim considers: (i) The two causes jointly determine the effect, "as when a car crash is said to be caused by the driver's careless braking and the icy condition of the road."[108] But in this case neither cause alone would be sufficient for the effect – rather, each is a partial cause of it. This is not in conflict with the exclusion principle, as it denies only the possibility of distinct *sufficient* causes of a single effect. Case (ii): the "two" causes are really one cause, hence clearly not independent from "each other". Case (iii): one cause is *synchronically sufficient* for the other.[109] This will be the case, as we have seen, when (for instance) a fluid heats up its container in virtue of a succession of molecular collisions with the container in which the molecules transfer their energy to it, and in virtue of its temperature (which, you may recall, is defined in terms of the *average*

---

[106] Although Kim himself names it the 'causal exclusion principle', the principle, rather unfortunately, has nothing at all to do with Kim's well-known causal exclusion argument, or the attendant problem of causal exclusion. I will say more about this at the end of this section, and examine the causal exclusion argument in detail in chapter 5.

[107] See his [1989a] for this statement of the principle, and the details of the argument that follows; and see also his [1998] pp.64-5 for an enumeration of the possible counterexamples. I have replaced Kim's usage of 'complete' with 'sufficient', due to my earlier appeal to 'complete' causes in defining $C_P$. The point that Kim appeals to the absence of counterexamples to justify the exclusion principle is due to Marras, and is well made in his [1998].

[108] Kim [1989b] p.254.

[109] Kim frames this case in terms of supervenience rather than sufficiency. See my discussion in 1.4 for details of the connection between these two relations.

molecular kinetic energy). These two causes, although distinct and each causally sufficient for the heating, are clearly not independent, for the aggregate of the molecules and their velocities is sufficient for the temperature of the gas, and the temperature of the gas is *nothing over and above* (and so depends on) the molecules and their velocities. Case (iv): there is a causal relationship between the two causes. In the double assassination case, as we have seen, the two bullets are effects of a common cause, hence not independent. Other causal relations are possible here – one cause might cause the other, for instance. This licenses two further possibilities. Either the two causes are successive links in a causal chain leading to the effect, or else one cause simultaneously causes the other, and both then cause the effect at the same time. In none of these cases are the two causes independent. Case (v): This is a case of genuine overdetermination. We have to understand 'genuine' here as 'coincidental'. I am not sure that Kim would agree; no matter, for my purposes, genuine overdetermination is the nasty sort that, if it happens at all, is extremely rare due to the absurdity of widespread coincidence. The 'unless' clause in the principle explicitly exempts it from application to such cases.

I do not know how to prove that there are no further cases, but I can't think of any. So, the principle as stated has no counterexamples. Notice, however, that in exempting itself from application to cases of 'genuine overdetermination', the principle (provided there are no further cases) is a tautology! That is, the manner in which we have understood the (admittedly rare) cases of genuine overdetermination *just is* in terms of two distinct, independent causes are simultaneously sufficient for a given effect. What's strange about the assassin case that cries out for explanation *just is* that the two appear to be independent, sufficient, simultaneous causes of the same effect. But now the principle says nothing. Kim, it seems, agrees: in his [2003], Kim acknowledges that the principle, stated in the above manner, is analytic! The analyticity is easily remedied, however, for the proposition *that genuine overdetermination is extremely rare* is not analytic. The conjunction of this proposition with the exclusion principle means we can replace the 'unless' clause with the usual '*ceteris paribus'*. But which of the cases is appropriate in the case of

apparent overdetermination by the mental and physical events that cause our bodies to move?

Of the five possibilities mooted, arguably only (ii) and (iii) will do the trick. Case (i) is ruled out by $C_P$, for that principle (which we are assuming true) tells us that all physical events have sufficient *physical* causes. Case (v) is ruled out, as the co-occurrence of mental and physical sufficient causes is not rare at all, but happens all the time. Case (iv), on the other hand, is more tricky. Suppose neurophysiological events *cause* mental events, and both these events then cause a given bodily movement. Since there is no principled limit to how small the interval between your making a decision and acting on that decision can be, it follows that if both these causes are to precede their effects, the neurophysiological event and the mental event must be *simultaneous*. It is difficult to rule out simultaneous causation, as I mentioned in 1.4. There, you will recall, I argued that synchronic causal sufficiency does not satisfy our definition of synchronic sufficiency. However, it is one thing to *distinguish* causal from non-causal synchronic sufficiency, and quite another to argue that the former does not occur. Here is an (admittedly inconclusive) argument to the effect that simultaneous causation is not appropriate as an explanation of the co-occurrence of mental and physical causes.

As I argued in 1.4, it must be possible for causes to fail to have their effects. All sorts of things can get in the way of a causal relation. So how come my behaviours *always have mental causes*? Of course, there are bodily movements, like twitches, that *lack* mental causes. But these are not behaviours. The point is this: if $C_P$ is true, then my behaviours will have physical causes. But if the relationship between the physical causes and the mental causes is *causal*, then how come we don't see more cases of psychologically *uncaused* bodily movements that nonetheless *look just like behaviours*? It should be possible, if the relationship between mental and physical is causal, for the physical causes of my behaviours to cause them without causing the mental causes, but we don't observe the sort of irrationality that this possibility licenses. So the evidence suggests that the relationship between mind and body can't

be causal. This argument is inconclusive, as I said, for nothing in it prevents a proponent of the simultaneous causation theory insisting that the causation of the mental by the physical is strict. Nonetheless, I will continue as if (iv) is not an option, given the widespread co-occurrence of mental and physical causes of behaviour. If these remarks are correct, then $O_D$ tells us that two events or properties causally sufficient for the same effect are either identical or related by (non-causal) sufficiency. Now as I defined sufficiency in 1.4, identity is a limiting case. We may therefore state $O_D$ as follows:

> $O_D$: if an event *y* has a sufficient cause *x* at *t*, then no event *x*' is a cause of *y* at *t* unless (i) *x* is sufficient for *x*', or (ii) *x*' is sufficient for *x*.

We include the qualifier 'at *t*' for obvious reasons: without it, the principle is false, for events will have many sufficient causes given the transitivity of causation. An obvious objection. As formulated, $O_D$ may seem unsupported by the arguments given. Those arguments, after all, surely only rule out independent causes *each of which is sufficient*? But in the above formulation, I make the stronger claim that if there is one sufficient cause of an effect, then *any* other cause of it can not be independent of the first cause. Surely, the objector continues, $O_D$ should be formulated like this:

> $O_D^2$: If an event *y* has a sufficient cause *x* at *t*, then no event *x*' is a sufficient cause of *y* at *t* unless (i) *x* is sufficient for *x*', or (ii) *x*' is sufficient for *x*.

It is with good reason that I prefer $O_D$ to $O_D^2$ – if the causal argument is formulated in terms of $O_D^2$, it faces obvious objections. The problem is that with $O_D^2$, $E_M$ must take up the slack – that is, we must now claim not just that mental events cause physical events, but that mental events are *sufficient* causes of physical events, for otherwise $O_D^2$ will not apply and the argument will be invalid. On the other hand, it is far from clear that $E_M$ strengthened in this way is true. I will not get drawn into arguing the

point either way. This is because in fact, the stronger $O_D$ is every bit as well supported by the arguments given as $O_D^2$ is. Allow me to explain.

The important thing to notice is that given $C_P$, *whatever* causal role the mental has will be such that unless the mental and physical causes are not independent, we are left with coincidences of an unacceptable sort. Consider again the two assassins, but this time suppose that one of them has a gun loaded with rubber bullets. The causal role of this assassin is that he makes the victim's death slightly more probable – there is a non-zero probability that the victim would die as a result of the impact of the rubber bullet in the absence of the other assassin's bullet. Crucially, we *still require an explanation* of why there are these two independent causes, despite the fact that they do not have the same causal potency with respect to the specified effect. The explanation might be that the assassins agreed beforehand that neither of them ought to know which one fired the fatal shot – firing squads are set up this way as a matter of course. The central point here is that the overdetermination problem that licenses $O_D$ does not consist in *both* causes being sufficient – rather, the worry is that given that *one* of the causes is sufficient, then *whatever* the other one is doing, unless there is an explanation of why it is doing it, we have a coincidence. We can make the point more dramatically: Suppose, just as the assassin's bullet strikes, a clown rides up to the victim on a monocycle and shouts "sausages". There is a non-zero probability that the victim (killed, we may assume, by the bullet) would die of a heart attack even in the absence of the bullet. If that sort of thing happens *all the time*, then it stands no less in need of explanation than more conventional constructions where the clown is replaced by another assassin. The fundamental things apply: if a sufficient cause of some effect is always accompanied by another cause of a different type, but a cause of the same effect nonetheless, then the two causes cannot, on pain of widespread coincidence, be independent. Of course, if they fail to be independent in virtue of being *identical*, then both causes will be sufficient. But in cases where one cause is causally sufficient for the effect and non-causally sufficient for the other cause, $O_D$ leaves open the possibility that the causal roles of the properties differ.

Before proceeding, I want to clear up a few misunderstandings that might arise. In the literature, it is a commonplace that distinct and unrelated notions of overdetermination are run together. Kim distinguishes the different notions, but often speaks as if they were more of less equivalent. In his [1993b] he asks us to consider a case in which a supervenient mental property M causes the instantiation of another supervenient mental property M*, where M is subvened by P, and M* by P*, and P causes P*.[110] Kim argues that the only way for M to cause M* is by causing P*, which we may grant him for now. Kim's asks 'whether M should be given a distinct causal role in this situation', and answers in the negative, thus:

> First, there is the good old principle of simplicity: we can make do with P as P*'s cause, so why bother with M?…Moreover, if we insist on M as a cause of P, we run afoul of another serious difficulty, "the problem of causal-explanatory exclusion". For we would be allowing two distinct sufficient causes…of a single event. This makes the situation look like one of causal overdetermination, which is absurd…The exclusion problem, then, is this: Given that P is a sufficient cause of P*, how could M *also* be a cause…of P*? What causal work is left over for M, or any other mental property, to do?[111]

There are three separate strands in this passage, which really ought to be kept apart. *The first strand* involves is an appeal to the familiar epistemic principle known as 'Ockham's Razor': that we should not multiply entities beyond what is necessary to account for the phenomena. But this is no mere 'principle of simplicity' – when Kim says we can 'make do with P as P*'s cause' he means that given that P is sufficient for P*, we don't need any further entities in order to account for the occurrence (/instantiation) of P*. Unless supervenient properties add something to the world that isn't already there in virtue of their subvening properties, then they are somehow redundant, and we may dispense with them. Something like this principle underpins just about *all* our theorising; however, the principle is very difficult to define. If $C_P$ is true, then physics will presumably account for all the phenomena; does that make all

[110] For Kim, there is no important difference between talking of events as causes, and talking of properties as causes, and he uses the two interchangeably. I do not think that anything much hangs on this terminology, provided it is understood that the properties in question are causes in the sense that they are the constitutive properties of efficacious *events*.
[111] See Kim [1993b] p.354.

properties not required to express completed physics redundant? In one sense, it does. We will look at redundancy arguments in chapter 6.2, where I will distinguish two ways ('singular' and 'general') in which a property can exhibit novelty, and argue that a property is redundant only if it is not novel in *either* sense. Given $C_P$, it follows (by the definitions I will give) that supervenient properties lack singular novelty; it is much harder to argue that they are redundant.

*The second strand* in the passage is the coincidence problem. Confusingly, Kim goes on to mention the 'problem of causal-explanatory exclusion', which he links to the *absurdity* of overdetermination. He does not say what he takes overdetermination to be, nor why it is absurd, but it seems clear it can't have anything to do with *explanation* – if it did, why would we need Ockham's Razor? Surely *that* principle is needed precisely because multiple causal explanations of the same phenomena *do* make sense. What rules out a second causal explanation of a given effect is not that, given the first explanation, it is absurd, but that it is (arguably) *otiose*. As we have seen, the absurdity of cases (like the one where two assassins shoot the same person) of overdetermination is that they seem to involve coincidences, which nobody likes. Of course, isolated coincidences are not absurd at all – if anything *is* absurd here, it is the supposition that such coincidences *happen all the time*. However, it is clearly no coincidence, in the example Kim describes, that two events that stand in a supervenience (sufficiency) relation occur at the same time. But now, obviously, there is nothing coincidental about the overdetermination that happens in the case of supervenient causation![112]

*The third strand* I want to distinguish in the quoted passage is the thought that "after" P has done its "causal work", there is nothing left over for M to do. The causal work

[112] This point is well made by several authors. For instance, Loewer [2001b] terms this sort of overdetermination 'dependent' as opposed to 'independent', and rightly maintains that only widespread overdetermination of the latter sort is problematic on grounds of widespread coincidence. Sider [2003] makes a similar distinction, but calls dependent overdetermination 'systematic'; he too rightly maintains that there is nothing absurd in supposing such overdetermination to be widespread. And Funkhouser [2002] draw the same distinction by distinguishing independent (coincidental) from 'incorporating' overdetermination.

to be done presumably involves something like pushing and pulling the various bits and pieces until they are so structured that P* is instantiated. If you want to build a wall, your work will involve cutting and moving bricks, mixing cement, laying bricks, and other tasks. If it overwhelmingly plausible that at least *some* cases of causation *involve* causal work – causing something involves making something happen, and if you want to do that, you have to be prepared to work at it. Now *if* causes are responsible for doing this causal work, then it does seem hugely plausible, given that P is sufficient for P*, that there isn't really anything left for M to do. For surely no part of the causal work involved in causation is done twice? This, as Kim says, is the causal exclusion problem. But notice, as I said above, that this problem has nothing at all to do with the causal exclusion *principle*, for as saw, that principle is concerned with coincidences and not causal work. We examine the causal exclusion problem in chapter 5.

As far as I can make out, just about every philosopher who has expressed a worry that the efficacy of mental or other properties seems to be 'screened off' or excluded by the physical has had in mind one or more of the three problems briefly characterised above. Sider agrees, and thinks that something very much like these three problems must represent the central concerns of those who worry about overdetermination.[113] Sider calls the three worries epistemic, coincidence and metaphysical, and they are exactly analogous to the three problems detailed above. There may something *else* that is bad about effects that have two causes, which may be the true ground of the worries about mental causation – Sider speaks of a "phantom fourth reason", but he, like me, is unable to see what it could be. Now it is possible to run arguments for physicalism based on any of the above conceptions of overdetermination, and Kim does exactly that. However, as Kim understands them at least, the arguments based on redundancy and causal exclusion are arguments for type identity. In fact (and the quoted passage is a case in point) Kim most often deploys these arguments not as arguments *for* type identity physicalism, but as arguments *against* supervenience

[113] In his [2003].

physicalism. Kim's strategy in deploying them is to start with the strongest form of non-identity relation between mental and physical properties (for which purpose he enlists his own strong supervenience) and show that even if this relation obtains, mental properties are either inefficacious or redundant. The exclusion argument purports to show if mental properties supervene on physical properties, there is no way to account for the efficacy of mental events, as there is no causal work left for them to do given their supervenience bases. Physical properties do all the causal work, and only properties identical to them are genuinely efficacious. This argument is the subject of chapter 5. The redundancy argument purports to show that if mental properties supervene on physical properties, then they do not have genuinely novel causal powers, and anything they causally explain can be explained just as well by their base properties. Only physical properties have genuine explanatory novelty, and any properties not identical to them will be redundant. We will examine this argument in 6.2. For my part, I think that the redundancy argument is much easier to resist than the exclusion argument – that's why the latter gets a full chapter, the former only a section. For the moment, notice that both these arguments, in pushing physicalism towards type identities, are the sorts of arguments most physicalists would spend their time *resisting* rather than defending. Multiple realization is the consensus view, and as we saw in 2.3, you can get into an awful mess if you try to combine multiple realization with type identity. Physicalism should be neutral between the various possible metaphysics of mind, and should certainly not entail eliminativism; as such, physicalism demands an argument that establishes supervenience rather than identity. That is why, in formulating the causal argument, we shall ignore the redundancy and exclusion problems, and see if the coincidence-based principle formulated above is enough to give us the kind supervenience we want.

### 3.4. How the causal argument works

Before we formulate the argument, a quick note is in order to resolve an issue that arises from the way in which I have defined $O_D$. Recall the definition that I gave above:

> $O_D$: If an event *y* has a sufficient cause *x* at *t*, then no event *x*' is a cause of *y* at *t* unless (i) *x* is sufficient for *x*', or (ii) *x*' is sufficient for *x*.

Now suppose *x*' is a mental event and *x* is a physical event. This definition leaves open the possibility that the mental cause is sufficient for the physical cause *without them being identical*. This is the position, of course, of idealists such as Berkeley. Berkeley thought that we perceive only ideas – we are never aware of anything outside our own mental lives. However, we clearly perceive ordinary objects – it follows that ordinary objects are ideas. It is not difficult to respond to this sort of argument, as it seems to clearly equivocate between direct and indirect perception. It is only plausible that we only perceive ideas if perception is understood in a direct, immediate sense. But then it is scarcely plausible that we are aware of ordinary objects in the same way. A similar position was once endorsed by Russell, who also thought that the traditional view of the relationship between the material and the mental got it backwards.[114] He thought this on epistemological grounds: (i) we know quite a lot about material objects, (ii) certainty is necessary for knowledge, (iii) we can only be certain of the objects of our acquaintance, i.e. 'sense-data', so (iv) knowledge about material objects is *really* knowledge about sense-data. It's true that Russell treats sense-data as objective, intersubjectively available entities; however, this is arguably just a mistake. To the extent that they are intersubjective, it is unclear how our knowledge of them can be any more certain than our knowledge of material objects, for I should then have to know about *everyone else's* sense-data in order to be acquainted with an object. However, I am acquainted only with *my own* sense data; from this it follows that I am not acquainted with material objects. But once this much

[114] Russell [1917].

is allowed, the epistemological motivations for positing sense-data evaporate, and we might as well stick with an ontology of mind-independent *objects*. In what follows I take it that the reasons for rejecting idealism, phenomenalism, and other doctrines that proclaim the ontological priority of the mental over the physical, are well understood. Having said that, it is far from clear to me that the majority of what I have to say would fail to apply, *mutatis mutandis*, if such a doctrine were true. Given the ontological priority of the physical, however, we may rule out the sufficiency of the mental for the physical. Call the priority claim $O_P$, and define it as the claim that unless the mental and physical events are identical, the mental ones are never sufficient for the physical ones.

We can now generate the following causal argument for physicalism:

$E_M$: Mental events cause physical events.

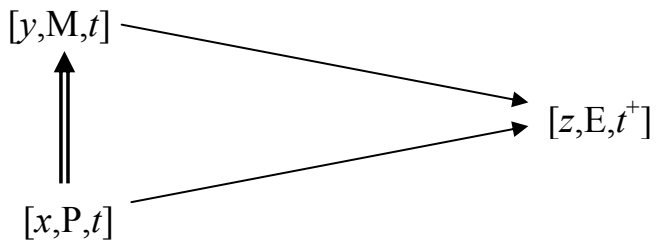
$C_P$: Every physical event *y* that has a sufficient cause at *t*, has a complete, sufficient *physical* cause *x* at *t*.

$O_D$: If an event *y* has a sufficient cause *x* at *t*, then no event *x*' is a cause of *y* at *t* unless (i) *x* is sufficient for *x*', or (ii) *x*' is sufficient for *x*.

$O_P$: Mental events are not sufficient for physical events unless mental and physical events are identical.

∴ Physical events are sufficient for mental events.

The situation can be depicted like this:




$E_M$ fills in the causal arrow from mental event [*y*,M,*t*] to physical effect [*z*,E,$t^+$]; $C_P$ fills in the arrow from physical event [*x*,P,*t*] to [*z*,E,$t^+$]; and $O_D$ fills in the sufficiency

arrow from to [$x$,P,$t$] to [$y$,M,$t$]. There are many diagrams similar to the one above in what follows; the convention I follow throughout is that single arrows indicate causation, and double (vertical) arrows indicate synchronic sufficiency. In addition, my diagrams will always depict relationships between *instances* of properties (i.e. events), and not the properties themselves.

Now we have established that physical events are sufficient for mental events. It is worth quickly going over again how this is supposed to establish physicalism. Recall our definitions of physicalism in 1.2 (P6), and of sufficiency in 1.4:

> Physicalism is true just in case:
>
> $$\forall u \forall \mathrm{M} \in M\{\mathrm{M}(u) \rightarrow \forall y \forall z[\mathrm{W}(z).\mathrm{I}(y,z).\mathrm{C}(y,u).\mathrm{P}^{\mathrm{w}^*}(z,\mathrm{w_a}) \rightarrow \mathrm{M}(y)]\}$$
>
> [$x$,P,t] is α-sufficient for [$y$,M,t] just in case:
>
> 1. $\exists w \exists x[\mathrm{W}^{\alpha}(w).\mathrm{I}(x,w).\mathrm{P}(x)]$
> 2. $\forall w \forall x \forall t\{[\mathrm{W}^{\alpha}(w).\mathrm{I}(x,w).\mathrm{P}(x)_{\mathrm{t}}] \rightarrow \exists y[x^{*}y.\mathrm{M}(y)_{\mathrm{t}}]\}$

*Prima facie*, they do not look too similar. However, as I explained in 1.4, and as we saw again in 2.2 in the context of functional reduction, establishing sufficiency will be enough to yield physicalist supervenience, provided $\mathrm{W}^{\alpha}(x)$ = '$x$ is a physically possible world'. The reasoning is relatively simple. If physical events are α-sufficient for mental events, then physical properties are α-sufficient for mental properties. Now if α-sufficiency is physical sufficiency, then follows immediately that a minimal physical duplicate of the actual world is a world in which my physically indistinguishable counterparts have the same mental properties as me. Since we can generalise the argument over any efficacious mental properties and any individuals, it follows that a minimal physical duplicate of the actual world is one in which all our counterparts have their efficacious mental properties. On the assumption that all actual mental events and properties are at least capable of causing physical events, physicalism as defined follows from the physical sufficiency of physical events for mental events. Before proceeding, a brief mention is in order of a possible problem

posed by *broad content* for this argument. [*x*,P,*t*] above is the cause of a behavioural effect; as such, it seems that P must be an intrinsic property, and not the sort of relational feature that externalists standardly hold to determine content. But if this is true, then [*x*,P,*t*] will not be sufficient for [*y*,M,*t*], as it does not include certain determinants of M's content. I will give two brief replies. *First*, externalists typically do not want their externalism to preclude $E_M$; as such, they typically endorse theories of causation according to which the extrinsic properties of an individual make a difference to its causal powers. But if this is true, then there is no reason why P should not be extrinsic, as well as M. *Second*, if causal efficacy is limited to intrinsic properties, then [*x*,P,*t*] will be sufficient for the efficacious part of [*y*,M,*t*]. The missing part (M's content) should not worry a physicalist, for if externalism is true, then this part will involve relations to the believer's physical or social environment. Adding in, say, causal covariance with a natural kind, to [*y*,M,*t*] adds to it only extra physical properties. This being so, we can rest content with establishing physicalism about M's efficacious part, and let broad content take care of itself.

The position we are in is this: *if* the causal argument establishes that physical events are physically sufficient for mental events, then it establishes physicalism. Still, there is a question mark over the validity of the argument. The trouble is that the theoretical work that sufficiency does in the causal argument consists in rendering the co-occurrence of mental and physical causes non-coincidental. Problematically for proponents of the causal argument, nothing forces the view that the sufficiency must be as strong as *physical* sufficiency in order to do this work. The central question, then, is this: is there a form of non-causal sufficiency consistent with the premises of the causal argument, but *in*consistent with physicalism? If there is, then the argument is invalid. Forms of sufficiency weaker than physical sufficiency are easy to come by. For instance, if certain forms of emergentism are true, then there are extra, non-physical, laws according to which physical properties are sufficient mental properties. Now if that is so, then in our definition of sufficiency $W^{\alpha}(x) = x$ is a nomologically possible world. From this, however, we *can't* infer physicalism, as minimal physical duplicates of nomologically possible worlds fail to duplicate the extra laws. However,

it is no easy task to define an emergentist position that is consistent with the premises of the causal argument. We take up this task in chapter 6. For now, notice that the causal argument establishes (at least) the falsity of Cartesian *substance* dualism. For if I am right about the nature of non-causal sufficiency, then the view that physical events are sufficient for mental events entails that the constitutive objects of mental events have only physical parts. Notice also that even if it turns out that we can only establish nomological sufficiency, we will still be able to infer that the mental is nothing over and above "the *natural*". We might define a predicate $N^{w-}(x,y)$ = *x* is a minimal natural duplicate of *y*. If we replace $P^{w-}(z,w_a)$ with $N^{w-}(z,w_a)$ in the above definition of physicalism, we get a position we could reasonably term 'metaphysical naturalism' – that the mental is nothing over and above "the natural". And this is a position we *can* infer from nomological sufficiency, as this form of sufficiency by definition ranges over all possible worlds with the same laws of nature as the actual world.

Before proceeding to examine weaker forms of sufficiency, however, there are two arguments that demand discussion. The first, which we discuss in the next chapter, is due to Scott Sturgeon, who argues that the causal argument doesn't establish *anything at all*, due to equivocation on the sense of 'physical'. Clearly if this is true, then there is no need to worry about whether or not the causal argument establishes a form of sufficiency from which we can infer *physicalism*; if Sturgeon is right, then *a fortiori* it does not. The second, due largely to Kim, but endorsed, in one way or another, by many others, is the 'causal exclusion argument' to the effect that the only position consistent (on reasonable assumption) with the premises of the causal argument is *identity*. If Kim is right, then again we have no need to worry about whether the form of sufficiency the causal argument establishes is one from which physicalism can be inferred, for any putatively non-physicalist forms of sufficiency will be ruled out, along with any physicalist forms of sufficiency other than identity. We discuss the exclusion problem is chapter 5. We will find neither Sturgeon's argument nor Kim's argument compelling, and return in chapter 6 to the question whether there is a non-physicalist form of sufficiency consistent with the premises of the causal argument.

# 4. Does the Causal Argument Equivocate?

Sturgeon argues that the causal argument equivocates on the on the sense of 'physical'.[115] $E_M$, he suggests, is true for *macro*physical effects; but $C_P$ is true only for *micro*physical effects. In order to prevent equivocation, Sturgeon claims, physicalists need to endorse causal 'transmission principles' to push causal efficacy around between the macro and micro levels. Without such 'pushing around' of causation, there is no competition between mental and physical causes. He goes on to argue that the principles are unsupported, due to the weirdness of quantum mechanics. But without causal competition, there is no causal argument, as our diagram of 3.4 suggests. In this chapter, we will first see why Sturgeon thinks the causal argument needs transmission principles, and formulate the principles needed. In 4.2, I examine Sturgeon's argument against transmission principles, and show that it fails. I find the argument obscure, and so am unconvinced that I have Sturgeon right. In 4.3, I give my own reasons for doubting causal transmission. My reasons are not conclusive, but they are good reasons for thinking that whether the principles are true or not, proponents of the causal argument would be well advised not to rely on them. And finally, in 4.4, we will see why, despite all this, the causal argument does not need the principles (or anything like them) in the first place.

## 4.1. Equivocation and transmission principles

Refer back to the causal diagram in 3.4. It depicts a mental and physical cause, competing for efficacy with respect to a *single physical event*. This is because $O_D$ is explicitly stated in terms of multiple distinct causes of a single effect. In 3.3, we motivated $O_D$ by reference to double assassination cases – without some form of collusion between the assassins, the fact that both assassins cause the death would be a coincidence. The way we have set up the argument, then, seems to require that the mental and physical causes are both efficacious with respect to the same effect.[116] But now, as Sturgeon notes, there is a problem. $E_M$ claims that mental events cause

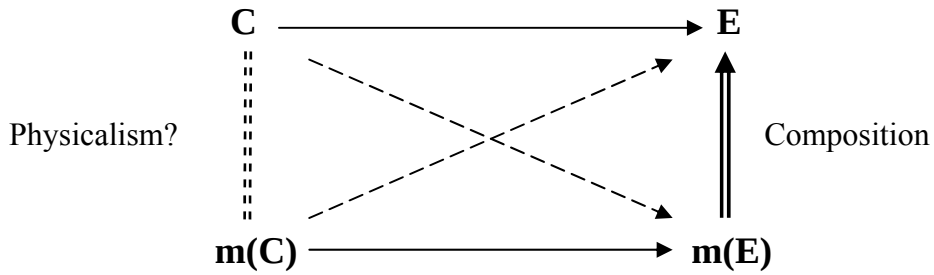
[115] Sturgeon [1998].
[116] This is exactly what I will deny in 4.4. I will also, correspondingly, offer a reformulation of $O_D$.

physical events, but the sense intended in here is that of 'broadly physical'. Mental events cause things like arm-movements, muscle contractions, and so on. $C_P$ claims that all physical events with sufficient causes have complete, sufficient physical causes. However, if we read the sense of 'physical' in $C_P$ as 'broadly physical', then it is unsupported. The evidence for $C_P$ does not justify the claim that all *broadly* physical effects have *broadly* physical causes. Indeed, it seems clear that some broadly physical events *lack* broadly physical causes – think for example of a nuclear explosion caused by the radioactive decay of a Uranium atom. The 'physics' in 'the completeness of physics' is PHYSICS, not *folk*-physics. And if the arguments of 3.2 are correct, and there is an inductive connection between current physics and PHYSICS, then it seems just as clear that future theory will not contain arm-movements as it does that future theory will not contain mental forces. Sturgeon goes further, and claims that it is the *basic* physical level that is complete, equating this level with quantum mechanics. I disagree on both counts, First, quantum mechanics is *not* complete. Second, there is no reason to suppose that PHYSICS will occupy a single level in the micro-macro hierarchy. The claim that PHYSICS won't involve macro events like bodily movements and earthquakes is one thing; the claim that it will involve only the entities of a 'fundamental micro level' goes well beyond anything the evidence might support.[117] No matter; the appeal to fundamentalism is not needed to motivate the problem Sturgeon raises. The central point is that if we read the sense of 'physical' in $E_M$ as PHYSICAL, then it too is unsupported by the phenomena: we observe constant conjunction between mental events and arm-movements, not accelerations of basic physical particulars, or fluctuations in electromagnetic fields. It is no part of either folk-psychology or folk-physics that the mind can influence the quantum world.

Whence a problem. There is no apparent causal competition, if mental events have broadly physical effects, but the domain of the broadly physical is incomplete; and if PHYSICS is causally complete, but the mind does not have PHYSICAL effects. An

[117] See Schaffer [2003] for convincing arguments to the effect that there is no current evidence in support of the view that a fundamental level even *exists*, let alone that a causally complete one does.

obvious response (the one Sturgeon wants us to give) is that the macrophysical and microphysical domains are metaphysically related to each other. There is a variety of relations to choose from. Big events like earthquakes are certainly *composed* of smaller events; and the small events taken together as aggregates are *sufficient for* the big events.[118] Now let C be the mental cause of some ordinary physical event E; and let m(E) be the microphysical cause of the microphysical events m(E) that compose E. We may depict the situation thus:




We want to establish the physicalism about C via *causal competition* between C and m(C) for some effect. But C causes a behaviour E, while m(C) causes not E but the microphysical events m(E) that compose E. This is what we need transmission principles for – they fill in the dotted causal arrows, generating the required competition. Sturgeon considers composition as a candidate relation.

> 1. Causal transmission under upward composition
>
> If an event m(C) causes an event m(E), and m(E) composes E, then m(C) causes E. Under this principle, if behaviours are composed of quantum events, and these latter have quantum causes, then so too do behaviours. This generates conflict with the efficacy of the mental with respect to behaviours.

[118] I omit supervenience in what follows for the reasons detailed in 1.4. I do not think that supervenience is appropriate to characterise noncausal relationships between particulars – rather, as I explained, I prefer to think of the matter in terms of sufficiency, which is of course intimately related to supervenience.

2. Causal transmission under downward composition

If an event C causes an event E, and E is composed by E*, then C causes E*. This principle has the consequence, again if behaviours are composed by quantum events, that some quantum events have mental events as causes. Once again we have the required competition.

Principle (1) fills in the dotted causal arrow from m(C) to E, in which case C and m(C) are in competition for E; principle (2) fills in the downwards arrow from C to m(E), bringing C and m(C) into competition for m(E). Sturgeon's strategy is to create problems for the principles, by arguing that (i) there are certain composition relations between m(E) and E such that the principles fail, and (ii) because there is a conceptual gap between basic microphysical (quantum) reality and macro (broadly physical) reality (largely due to the strangeness of quantum physical stuff compared to broadly physical stuff), *we don't know* whether the relation that holds between basic physics and behaviour is of the right sort. Therefore, we don't know whether the principles are true in the required cases, and so we shouldn't endorse the causal argument.

Notice at this point that this argument need not trouble proponents of *non-mentalism*. This is because there is a range of uncontroversially non-mental effects that can take the place of E in the above diagram. For instance, it *is* part of our common-sense world view that mental events cause houses to be built. But houses are clearly non-mental events. So if you think that the *non-mental* is complete (as you ought to if you accept the non-triviality argument of section 3.2) then the non-mental cause of E will be non-causally sufficient for C.[119] And from this we can infer non-mentalism about the mind, without having to worry about what goes on in any micro domain. As I argued in 3.2, however, non-mentalism is a poor substitute for physicalism proper. If this much is granted, then a reply to Sturgeon is clearly preferable to abandoning physicalism in favour of non-mentalism. In the next section, we will consider what I

[119] As such, nothing in Sturgeon [1998] will cause problems for Spurrett and Papineau [1999].

take to be Sturgeon's argument against the transmission principles, and find that it fails.

### 4.2. Sturgeon's argument against transmission

Sturgeon points out that certain constraints must be placed on the so far undefined notion of 'composition' if the transmission principles are to come out true. For the sake of argument, Sturgeon says, let the composition relation in question be that of partial mereological constitution – some parts of the behaviour are quantum events. Imagine someone becomes hungry and so grasps an apple; at the same time, a random chemical reaction occurs in their brain that causes their little finger to twitch. The twitch by the present definition composes the grasp, and yet the chemical reaction does not cause the grasp, hunger does. Hence a counterexample to (1). And according to the present definition, the grasp is composed by the twitch, and yet hunger does not cause the twitch, the chemical reaction does. Hence a counterexample to (2). Partial mereological constitution won't do, but why not? Intuitively, the answer seems to be that the twitch is somehow *inessential* to the grasp, and Sturgeon follows this intuition to the 'Cause and Essence Principle', which states that:

> C causes E iff C is sufficient to bring about what is essential to E.

This principle will clearly place constraints on the kind of composition relations that will satisfy the transmission principles. It rules out partial mereological constitution for obvious reasons – in the upwards case, for instance, the chemical reaction fails to cause the grasp just because it is not sufficient to bring about what is essential to the grasp. The only candidate it causes is a twitch, which is 'inessential'.

It is unclear to me exactly what work the appeal to essentialism is supposed to be doing in Sturgeon's argument; what follows is at best a reconstruction, one that I hope comes close to what Sturgeon intends. In general, let us grant that token events whose parts are other events possess some of those parts essentially, others inessentially. What is essential to an arm-movement, say? Sturgeon claims we know *a priori* that

'[t]he essence of hand movements consists in sub-hand movements.'[120] This is true, strictly speaking, only under the assumption that everyday things are composed of parts, and that's an empirical matter. Still, that arm movements are composed of arm-part movements is somehow less empirical than that that they are composed of muscle-and-bone movements, say. That represents a theory about what arm parts actually *are*. There's a *conceptual gap* between muscle-talk and arm-talk, which is closed by a combination of (relatively-conceptual) analysis and empirical theory. Now this is very close to the functional reduction procedure we examined in chapter 2. What matters here – what closes the conceptual gap – is a theory that explains how it is that muscles and bones get to play the causal roles that characterise arm parts. So: (i) it's essential to an arm movement that the parts of an arm move; (ii) in the actual world, these parts are muscles and bones and other stuff that I don't know about. Now Sturgeon seems to want to infer from (i) and (ii) that token muscle and bone movements are essential to (actual) token arm movements. This seems right – if it is true that all actual arm-parts are muscles and bones, then it will be true of actual token arm movements that they could not have occurred without muscle and bone movements. In the case of the grasp described above, I think we may say with some confidence that the very same token grasp could have occurred without the twitch. But the grasp is also composed of muscle and bone movements, and however robust it is, it will presumably not be so robust that it (this very same token grasp) could have occurred without any of the token muscle and bone movements that actually compose it. The central point is this: the reduction of arm-movements to physiology tells us the essence of arm-movements. And this in turn means we can be confident about pushing causation around between muscle movements and arm-movements. However, Sturgeon claims, the same is not true in the case of the relationship between quantum events and macro events in general. Why not?

Sturgeon's argument, in a nutshell, is this: since there is a "yawning conceptual gap" between quantum reality and macro reality, we do not know whether or not the

[120] Sturgeon [1998] p.422

composition relation that obtains between these domains is such as to satisfy the cause and essence principle. But it is unclear what particular gap in our knowledge of the quantum world is driving the doubt. Presumably we have already finished the *a priori* part that Sturgeon envisages for conceptual gap-closing; let's say that macro object parts are essential to macro objects. What we lack in the case of quantum mechanics, I suggest, is the empirical bit that explains how it is that quantum events compose macro events. Sturgeon seems to be worrying about the 'measurement problem' here. In a (very small) nutshell, here is how it goes. Quantum theory has two elements, which seem to be in conflict with each other. They are: (i) equations that describe the dynamic evolution of quantum systems, and (ii) the 'collapse postulate' according to which some of the properties ascribed in (i) disappear when a measurement is made. Now (i) has enjoyed huge predictive success, and (ii) is *ad hoc*. But (ii) is needed precisely because the properties ascribed in (i) are not the sort of properties we ever observe. For instance, (i) ascribes 'superposition states' to quantum particles, in which – to a degree specified in the theory – they have both of two incompatible properties, such as 'spin-up' and 'spin-down'. When we measure them, however, we find that they have – determinately – one property or the other. What is really odd is that the degrees assigned in (i) turn out to be highly accurate predictive indicators of the frequency with which we observe each property. If we take the mathematical formalism of quantum mechanics at face value, it entails that when you measure whether a superposed particle is spin-up or spin-down, the measurement system, including experimenter and apparatus, become 'entangled' with the quantum state. That is, we get a new quantum system – an 'experimenter-apparatus-particle' system – which is a superposition state of the two classical systems 'experimenter-observing-spin-up' and 'experimenter-observing-spin-down'. Needless to say, this theory does not sit well with what experience tells us about such situations. The collapse postulate says that something happens when a measurement is made that forces the quantum world to give up its inherent oddness and fit in with the macro stuff, according to which the particle has one spin or the other, and to no degree both. The trouble is that

there is no explanation of what is so special about measurement, compared to the other interactions that quantum particles take part in.[121] Why on earth should quantum particles behave like classical particles just when they are being measured, if quantum theory is right that they behave non-classically at all other times? Suffice it to say that there is no agreed solution to this problem. This, I take it, is pretty much what Sturgeon has in mind when he speaks of 'yawning conceptual gaps' between quantum and macro reality.

Why is any of this a problem for causal transmission? Sturgeon's reply to Noordhof is illuminating.[122] Here's an example:

> It's a fact of life that macro movements are composed of quantum phenomena. The conceptual gap precludes knowing whether the latter are essential to the former. By the Cause and Essence principle, we should resist [the upwards transmission of causation from quantum to macro events]. *Since it's unclear quantum phenomena are essential to macro movements*, it's unclear causes of the former thereby cause the latter. [Italics mine.]

This is an epistemic point: the lack of a theory that explains *how* quantum events compose macro events (given the apparently inconsistent properties of quantum and macro phenomena, or so I assume) leaves open the possibility that quantum phenomena are to macro phenomena as twitches are to arm-movements. I am happy to grant Sturgeon his essentialist claims; if I seem to do so too freely, then it is only because there are completely irrelevant to his conclusion. A brief argument is in order to explain why. Suppose for the sake of argument that all members of a certain class of events – the class of bodily movements, say, for consistency of exposition – are maximally robust. By maximal robustness, I intend that such events do not have *any*

[121] Some authors claim (as seems to me to be right) that there is nothing special about measurement, and that the collapse postulate is false. The burden of such a view is to explain away the apparent truth that the macro world does not contain superpositions of live and dead cats. See Papineau [1996] for such an explanation. Others, such as Wigner [1962] claim that interaction with conscious minds has an effect on quantum states that no quantum state can have – quantum physics, on such a view, is causally incomplete with respect to phenomenology. This view has few adherents. See Hughes [1992] (especially chapters 9 and 10) for discussion of how the measurement problem relates to the problem of reconciling the macro image with the odd properties of the quantum world.

[122] The passage below is quoted from Sturgeon's [1999] response to Noordhof's [1999a] reply to Sturgeon [1998].

of their proper parts essentially. The token events that constitute my typing this paper, we will say, occur just the same at a world where my finger movements are composed not by muscles, bones and suchlike, but by jelly. It follows that at *this* world, muscles and bones ought to fail the test supplied by the cause and essence principle, as by hypothesis muscle and bone movements are not essential to arm movements. But in fact they pass the test. Why? Because causing an m(E) that is *not* essential to E *is* "sufficient to bring about what is essential to E", provided m(E) is synchronically sufficient for E. If, at the actual world, muscle and bone movements are synchronically sufficient for arm-movements, then causing muscle and bone movements is sufficient to bring it about that arms move. *It does not matter whether they are essential*. Upshot: if you want to cause a composite event by causing the events that compose it, don't waste your time worrying whether the events you cause are *essential*; worry instead whether they are *sufficient*. Consider the following transmission principle:

> 3. Causal transmission under synchronic sufficiency
>
> If an event m(C) causes an event m(E), and m(E) is synchronically sufficient for E, then m(C) causes E.

Now the cause and essence principle, far from generating problems for (3), actually entails it! For m(C) is causally sufficient for m(E), and m(E) is non-causally sufficient for E; E could hardly occur without "what is essential to it" and so m(C) is sufficient to bring about what is essential to E. But from this it follows, given the cause and essence principle, that the cause of m(E), in causing it, thereby causes E.

The question remains, *are* quantum events sufficient for macro physical events? If they are, then it looks as though we can generate causal competition via something like (3). Witmer responds to Sturgeon along very similar lines. Here is the core of Witmer's argument.[123] Witmer argues that the *downwards* transmission principle (2)

[123] See his [2000].

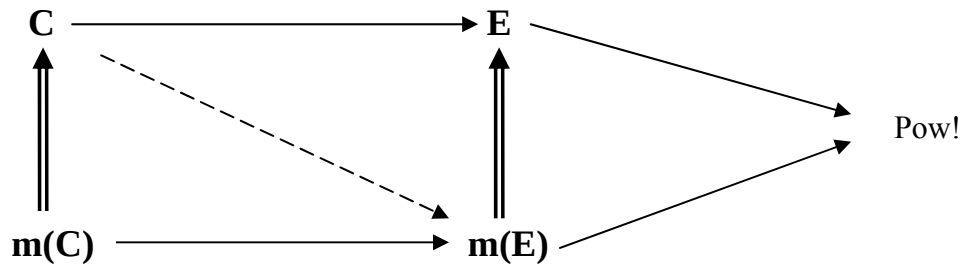
is true under supervenience relations strong enough to license the 'nothing over and above'-ness of the supervenient on the subvenient. His argument is that there is no way to cause dependent events like E other than by causing the events on which they depend, in our diagram, m(E). In other words, if E is nothing over and above m(E), how else could you cause E but by causing m(E)? Let us call this the 'downwards transmission argument', and let us grant Witmer the downwards transmission principle for the sake of argument.[124] Let us also recast the matter in terms of sufficiency, rather than supervenience – this is a terminological matter, for sufficiency, as I have argued, is a form of 'nothing-over-and-aboveness'. How are we to establish that m(E) is sufficient for E? Witmer appeals to our ability to *intervene* in the quantum world. He claims, in particular, that

> [O]ur knowledge of the theoretical-physical depends on experimentation. Passive observation is not enough; the sorts of hypotheses we test...are not liable to confirmed by [merely] observing the natural environment. Our knowledge...depends on an ability to *manipulate* the environment, including...the way the world is theoretical-physically.[125] [Italics mine.]

Put simply, we could not gain knowledge of the quantum world if we were not able to test hypotheses about the way the quantum world works, and in order to do that we need, *inter alia*, to set up initial quantum conditions and see if they evolve according to theory. Only the second part of this process is 'mere observation'; the first, by its nature, involves manipulation. This seems correct, and if it is, then at least some ordinary physical events have quantum effects. Of course, we can't manipulate the quantum world directly, but we can build machines to act as go-betweens. The clever thing about these machines is that they have buttons on them, which, when pressed, initiate causal chains that result in events like subatomic particles crashing into one another. Let such an event be 'Pow!'. We can depict the situation like this:

---

[124] In the next section we shall see that there are good reasons to resist transmission principles even if E is nothing over and above m(E). We return to the downwards transmission argument in our discussion of Kim's 'supervenience argument' in 5.1. Kim appeals to downwards transmission to show that the causal efficacy of supervenient properties is inconsistent with the completeness of physics. The burden of chapter 5 is to show that his argument rests on a theory of causation that is demonstrably false.
[125] Witmer [2000] p.284.




Causal competition for Pow! establishes (as in 3.4, on pain of coincidence) that m(E) is sufficient for (E). But if Witmer is right that the only way to cause a dependent event is to cause the events on which it depends, then we have further causal competition for m(E), which establishes, *mutatis mutandis*, that m(C) is sufficient for C. In the next section we shall see that there are good general grounds for avoiding the use of transmission principles if at all possible. In 4.4, we will see that there are two equally good reasons why the causal argument does not need them.

### 4.3. Counterfactual theories of causation

The upshot of 4.2 is that if our transmission principles fail, then it is thoroughly unclear what their failure has to do with conceptual gaps or the peculiarity of things quantum mechanical. At the end of his [1998], however, Sturgeon makes some remarks on causation and counterfactual dependency that are much more to the point.[126] There he says that microphysical events and ordinary physical events may exhibit different patterns of "counterfactual activity". Although actual microphysical events compose actual behaviours, they might come apart across possible worlds. Perhaps there are worlds at which *this very same token behaviour* occurs without the particular physical events that compose it at the actual world. But if causation *depends* on such patterns, then perhaps we do need to be careful about inter-level causal claims. The argument of this section turns Sturgeon's 'Cause and Essence' argument on its head: Sturgeon's problem for the transmission principles (as I understand it) is that we do not know whether quantum events *are* essential to macro events; the problem I raise will be that there is good reason to believe that at least some of them

[126] pp.428-30.

are *not*. I will then argue that *if* certain counterfactual theories of causation are true, then differences in the essential properties of metaphysically related effects entails the falsity of the transmission principles. I do not endorse the counterfactual theories discussed.

In what follows, we will consider the strongest form of non-identity we have to hand, viz. metaphysically necessary synchronic sufficiency. If the transmission principles fail on so strong a relation as that, then they are plausibly in some trouble. For surely, if there is to be any transmission of causal efficacy at all, it will occur in cases where m(C) causes an event m(E) which is, by hypothesis fully sufficient for E, thereby causing E? As I suggested in 1.4, aggregate events such as m(E) have different modal properties to the events they are sufficient *for*. If you object to aggregates then, as before, think in terms of pluralities – the same points will apply. It seems scarcely deniable that m(E) will be more modally fragile than E. If m(E) is the sort of aggregate I defined in 1.4, then it is as fragile as events get; perhaps there are aggregates that are more robust than that. However, I maintain that if you start changing the components of m(E) around (perhaps while keeping the same components, and simply changing their relations to each other), you lose m(E) before you lose E. Or in terms of plural quantification: not all of *those* events (which are the components of m(E)) are necessary for the occurrence of E. But *all* of those events are necessary for the occurrence of *those events*! This is just to repeat the familiar point that events like behaviours are relatively robust, in the sense that they could have occurred in a different manner and yet remained the same event. Grant this relative robustness and fragility for the sake of argument. Notice that nothing in this picture depends on any conceptual gap between m(E) and E, nor on anything specific to the nature of either. All we have assumed so far is a difference in modal properties across a sufficiency relation, which I take it is not particularly groundbreaking.

But now suppose further that some variant of Lewis' counterfactual theory of causation is true. Lewis defines 'causal dependency' as follows:

An event *y* causally depends on an event *x* iff:

1. if *x* had occurred, then *y* would have occurred;
2. if *x* had not occurred, then *y* would not have occurred.

Causal dependency, for Lewis, is a sufficient, but not necessary, condition for causation. A causal chain is defined as a sequence of events in which each event causally depends on the previous one; events *c* and *e* are related as cause and effect just in case there is a causal chain from *c* to *e*. Let us refer to condition (1) as expressing the 'positive dependency', and condition (2) the 'negative dependency' of effect on cause. Now this theory, as is well known, is not without its problems; indeed, Lewis himself eventually gave up on this particular version of the analysis.[127] My purpose, however, is not to defend the theory, but to show that if it is true, then the transmission principles are not.[128] Lewis holds that the counterfactuals must be evaluated in the following way: they are true just in case it takes a greater departure from actuality to make the antecedents true and the consequents false than it does to make the antecedents true and the consequents true. It is often said that the first condition is vacuous, as the closest world to actuality at which its antecedent is true just is the actual world. It is for this reason that I diverge slightly from Lewis in the way I prefer to evaluate the truth of such counterfactuals. I think not in terms of a closest world, but in terms of a *neighbourhood* of closest worlds.[129] Suppose I water my grass, causing the grass to grow. If (1) is vacuous, then the relationship between me watering the grass is causal provided the closest world in which I do not water the grass is one in which the grass does not grow. But this pattern of dependency is, by my reckoning, insufficient. Surely we also want it to be the case that if my watering the grass is the cause of its growing, then it would have grown in worlds where I used

[127] See Lewis [1973] for the original theory, and [2000] for the theory he adopted in its place. In the latter paper, he treats causation as a relation that holds between robust events just in case an *influencing* relation holds between the (definitionally) fragile *alterations* of those events consistent with their still occurring, but in a different manner. I will say more about this theory in 5.5.

[128] I will not attempt to defend counterfactual analyses against the many objections that have been raised, although I will return to the matter, in the context of evaluating the relative merits of probability and process accounts of causation, in 5.5.

[129] A similar suggestion as to the evaluation of counterfactuals is to be found in Nozick's analysis of knowledge as 'truth-tracking'. See his [1981] pp.167-96.

hard water instead of soft water, or a different colour watering can, or whistled while I worked. Condition (1) can incorporate such cases, provided (i) its truth is indexed to a neighbourhood of not-too-distant worlds, and (ii) *x* is relatively robust, and occurs just the same in a neighbourhood of worlds at which the manner of its occurrence is different. With this in mind, let us evaluate the causes in the diagram to see if counterfactual dependency holds *diagonally*, assuming that it holds horizontally.

Do the transmission principles hold on the theory of causation outlined above? In order to show that they do not, I will suppose, for reasons of simplicity, that the causes and effects in our diagram are *proximal* – in this case, causal dependency is *necessary* (as well as sufficient) for causation.[130] If causal dependency fails for diagonally related proximal events, so then does causation. Consider first the downwards case – let us see whether, on the counterfactual theory, M causes m(E). Clearly m(E) negatively depends on M for in the closest worlds at which M does not occur, neither does m(C). But m(E) by hypothesis both positively and negatively depends on m(C), and so it negatively depends on M as well. So far so good, but m(E) does not positively depend on M. For the set of closest worlds at which M occurs will contain worlds where it occurs with a supervenience base other than m(C). This is due to the (assumed) differential robustness of M and m(C), and our agreed method of evaluating positive dependency against neighbourhoods of closest worlds. But since m(E) negatively depends on m(C), it follows that m(E) does not positively depend on M. Among the set of closest worlds we use to evaluate the truth of (1) with respect to M and m(E) will be worlds at which M occurs but m(E) does not. Similar arguments show, *mutatis mutandis*, that upwards transmission fails – m(C) does not cause E, according to the counterfactual theory. The set of closest worlds at which m(C) occurs are, on reasonable assumption, worlds at which m(E) occurs as well. Again if aggregates are fragile, then all worlds at which m(C) occurs will be worlds at which it occurs in the same way. But m(E) is by hypothesis metaphysically sufficient for E,

[130] This is because a chain of stepwise depdenceny between temporally contiguous events *x* and *y* will only obtain if *y* causally depends on *x*. I think it clear that the arguments to follow could be reformulated without the proximality assumption, to apply in the general case. I will not attempt to do so here, however, for nothing in my present purpose requires such a detailed analysis.

and so at all the relevant worlds E occurs as well. So E positively depends on m(C). However, *negative* dependency this time fails. For the closest worlds at which m(C) *fails* to occur will be worlds at which it fails to occur in virtue of a very similar aggregate occurring instead, call it m(C)'. To see this, recall that aggregates are artificial, and m(C) is really many events. Worlds at which *most* of those events occur are *ipso facto* closer to actuality than worlds at which *none* of them occurs. And the closest of these worlds to actuality will be worlds at which m(C)' causes m(E)', another realizer of E. If such worlds are possible (and it seems difficult to deny) then they are closer to actuality than any at which E fails to occur. So E does not negatively depend on m(C).

Apart from objections directed against counterfactual theories of causation generally (which, as I have said, are irrelevant to my present purpose), I anticipate the following objection to this account. Proponents of the counterfactual theory will object to my somewhat loose method of evaluating the truth of (1). I reply that (1) is vacuous unless evaluated against neighbourhoods of worlds, and so plays no role in a counterfactual theory of causation operating on such strict criteria concerning which worlds are relevant to the truth of counterfactual claims. The response may now be that (1) is otiose, and (2) is the essence of the counterfactual theory; I reply that I have already given my reasons for thinking (1) is important. Even if I am not granted this much, upwards transmission will still fail, for its failure depends only on (2). I will not dwell on these matters here, for there are other theories of causation according to which transmission principles fail. In the remainder of this section, we will look at one of them; we consider others in 5.4.

Yablo endorses a counterfactual theory of causation according to which it is a necessary condition on causation between events that cause and effect are *proportional* to each other.[131] Proportionality of cause and effect, for Yablo, is a matter of satisfying the counterfactuals (1) and (2) above, and in addition the cause

[131] See his [1992] for details.

must be both *enough* and *required* for the effect.[132] Yablo gives the following counterfactual definition of *causal requirement* for properties related as determinate and determinable:

> A property **P** of an event *x* is causally required for an event *y* iff for all $\mathbf{P}^- < \mathbf{P}$, if *x* had been $\mathbf{P}^-$ without being **P**, then *y* would not have occurred.

Where $\mathbf{P}^-$ is determined by **P**. Correspondingly, Yablo says that a property **P** is *enough* for some effect *y* if, given that *x* has **P**, some $\mathbf{P}^+ > \mathbf{P}$ is not required, where '<' and '>' mean, respectively, less and more determinate than. Now consider properties in a determination relationship, such as redness and colouredness. Redness will be causally required for a given effect just in case the cause would not have had the effect if it had been coloured without being red; and conversely, colouredness will be enough just in case no property of greater determinacy (no specific colour) is required. I do not wish to endorse the thesis that mental and physical properties are related as determinate and determinable, but I do think that Yablo's central idea can be generalised to cover sufficiency. I think this because what is central is *that the determinable events can occur in a different way*, and not that the *explanation* of this fact is that they *are* determinable. Now provided events that stand in sufficiency relations differ in their relative robustness and fragility, it follows that he events that are 'sufficed for' can occur without the sufficient events. Refer back to the diagram in 4.1, and assume that C and E are modally more robust than m(C) and m(E) respectively. From this difference in their modal properties, it follows that C can occur without m(C), and E without m(E). Let us say that m(C) is required for an effect *x* just in case if C had happened without m(C), then *x* would not have occurred. And

[132] It should be noted that Yablo in fact rejects (1) in favour of the following, which he terms 'adequacy': 'if *x* had not occurred, then if it had, *y* would have occurred as well'. He does so specifically to avoid the alleged triviality of (1). For my part, I think that despite Yablo's protestations to the contrary (see n.57 Yablo [1992] p.), if (1) is trivial then so is his adequacy criterion. Yablo rightly says that a cause *x* will be adequate for its effect *y* "just in case *y* occurs in the nearest *x*-containing world *w* to the nearest *x*-omitting world *v* to actuality". But if the only difference between *v* and actuality is that *x* does not occur, then the nearest *x*-containing world to *v* will be the actual world, at which both *x* and *y* occur by hypothesis. Much better, in my view, is to avoid the alleged triviality of (1) by relaxing the criteria by which the truths of counterfactual claims are evaluated, and allowing a few extra worlds in.

let us say that C is enough for an effect *x* just in case if C had happened without m(C), then *x* would have occurred anyway. Now clearly, on these definitions, m(C) is not required for E, and C is not enough for m(E).[133]

So, Yablo claims, mental properties are sometimes better candidate causes for given effects than their more determinate physical realizers, since there are effects (like behaviours, which are robust relative to the precise manner of their occurrence) that *require* the mental cause but not the physical realizer. And mental events, correspondingly, won't be enough for certain effects (such as the precise manner of a behaviour), as more determinate, physical, events are required. Now it seems clear that commensurability of this kind will obtain primarily between events at the same level. Causation, for Yablo, is not "pushed around" at all – not surprising, as we have already seen that modal properties do not happily travel up and down interlevel relations like sufficiency.[134] My point in this section has not been to defend the theories of causation I have outlined. Rather, I outline them simply as theories according to which causal transmission principles come out false. There are undoubtedly other such theories. In fact, I am prepared to guess that any theory according to which the obtaining of causal relations between events involves their modal properties, is likely to be a theory that has trouble transmitting causation. Whether or not any such theory is true is, it goes without saying, a matter of considerable controversy. But that isn't the point. The point is that the causal argument loses much of its force if relativised to a specific theory of causation, or

[133] Notice that a similar moral can be drawn from the subset theory of realization discussed in 2.4, which, as I mentioned there, is similar in many ways to Yablo's theory. If the causal powers of C are only a proper subset of the causal powers of m(C), then the powers of C will be a subset of those required to cause m(E), in which case downwards transmission fails. I return briefly to this point in my evaluation of Kim's redundancy argument in 6.2, which relies on downwards transmission. I do not claim that the subset theory entails too that upwards transmission fails – the causal powers of m(C) will be a superset of the powers required to cause E. Without an extra condition such as Yablo's stipulation that causes must be required for their effects, on the subset account m(C) causes E *a fortiori*, so to speak.

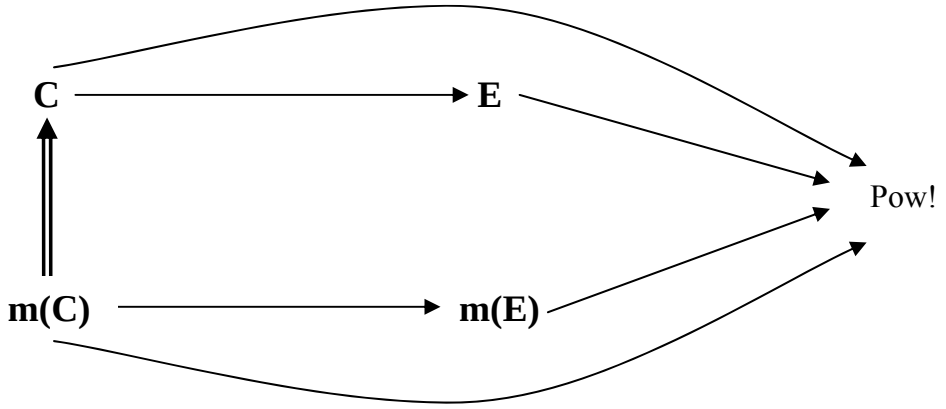
[134] Yablo is not alone in thinking that causation is intra-level rather than inter-level; Horgan thinks so too. In his [2001] he argues that causal claims have an implicit 'level-parameter'. Menzies [2003] argues, in a similar vein, that causal relations are relative to causal models, which he thinks of as particular systems of explanatory laws. Views such as these may be broadly categorised as 'compatibilist' and allow that there can be many non-exclusive (and perhaps complementary) levels of causation. We return to these issues in our discussion of the causal exclusion argument in chapter 5.

family of such theories. If the causal argument is to be resisted, then resisting it should be harder than simply adopting a counterfactual theory of causation.

Before proceeding, it is worth mentioning that I do recognise that the transmission principles have significant intuitive plausibility. There are several examples in the literature (most notably in Kim's work) of arguments similar to Witmer's to the effect that causing instances of supervenient properties just has to involve causing instances of the properties they supervene on. And if you manage to do that, then how (given sufficiency) could you thereby *fail* to cause the supervenient property-instance? My point in this section, as I said, has not been to endorse any theory of causation, but to show that the truth of certain such theories would entail that these causal intuitions just can't be right. In chapter 5 I will argue (*inter alia*) that the division between theories of causation that are, and are not, consistent with causal transmission, broadly tracks the familiar division of theories into *process* and *probability* accounts. The remainder of this chapter shows that the causal argument properly conceived is orthogonal to these issues, and requires no particular causal commitments. If I am right, then clearly that is a virtue of the argument.

### 4.4. The causal argument does not need transmission

The first reason that the causal argument does not need to appeal to dubious principles of transmission is implicit in Witmer's reply to Sturgeon. Recall that Witmer claims that unless at least some broadly physical events had quantum effects, we could not intervene in the quantum world, and so would not have epistemic access to it. In essence, Witmer thinks (and I agree) that certain events like button-pushings must have quantum effects. But he ignores the fact that the very same events must also have *mental causes* – it would hardly do us any good as interveners in the quantum world if the events that had quantum effects were events over which we had no rational control! If we are to test quantum theory by setting up quantum systems with certain initial conditions and observing how they evolve, then we must be able (indirectly, of course) to decide on the quantum properties of certain parts of the world. But if this is so, then we can draw the following diagram:




In this diagram, my decision C to intervene in the quantum world causes my pushing of the button, E, which in turn causes Pow! Now if C is capable of initiating a causal chain that ends in a quantum event which, by $C_P$, has a sufficient physical cause, it follows that m(C) must have occurred at the same time as C and initiated the same causal chain. Were this not the case, it would be possible to trace the causal ancestry of Pow! back to a point (the point where I decide to push the button) at which it had a mental cause but no physical cause. Transitivity fills in the arrows from C and m(C) straight to Pow! without having to worry about transmission or the relationship between E and m(E). The argument thus far is *ad hominem*: transitivity is implicit in the story Witmer tells about intervention, but if you have transitivity, then you don't need transmission. It is another question whether causation really *is* transitive.

Counterexamples to transitivity have been stacking up lately. Here are two of them.

Counterexample 1

*A* and *B* each have a switch in front of them, which they can switch up or down. If both switches are in the same position, person *C* receives a shock. *A* and *B* differ in that *A* does not want to shock *C*, whereas *B* does. Now suppose *B*'s switch is up. Since *A* does not want to shock *C*, *A* will move his switch down. But when *B* observes that *A*'s switch is down, she moves her switch down, and *C* receives a shock. It seems clear that *A* moving his switch down caused *B* to move her switch down, which in turn caused *C* to receive a shock. But did *A*'s moving his switch down cause *C* to

receive a shock? I have no clear intuition regarding this matter, but if you answer 'no', then for you this situation represents a failure of transitivity.[135]

Counterexample 2

A man walking in the mountains ducks to avoid a falling boulder. The man's ducking causes the continuation of his walk, by 'double prevention': the ducking prevents an event (the boulder striking him) that would have prevented the continuation of his walk. But the boulder caused him to duck. And yet it makes little sense to say the boulder caused the continuation of his walk – after all, it nearly killed him! My intuitions are clearer in this case, and I take it to be a failure of transitivity.[136]

Lewis claims that counterexamples such as these have a common structure. We are to:

> [i]magine a conflict between Black and Red…Black makes a move that, if not countered, would have advanced his cause. Red responds with an effective countermove, which gives Red the victory. Black's move causes Red's countermove, Red's countermove causes the victory. But does Black's move cause Red's victory? Sometimes, it seems not.[137]

Lewis defends transitivity via the claim that reluctance to accept Black's move as a cause of Red's victory stems from a conflation of what causes what with what is generally conducive to what. Moves like Black's are not generally conducive to the opponent's victory, but it does not follow that in this case, Black's move does not cause Red's victory. I am not sure what to say, and Lewis too admits to "feeling some ambivalence". Let us accept, then, that transitivity sometimes fails. However, it most definitely does not *always* fail. If it always fails, then not only do mental events not cause behaviours, barely anything that we think of as the cause of a given effect really causes it and almost everything we think about causation is false. The ways in which we identify certain events and properties as efficacious seldom picks out *proximal* causes, which is to say that there are almost always intermediaries between causes and

[135] This is due to Michael McDermott, see his [1995] for details.
[136] Hall [2000].
[137] Lewis [2000] p196.

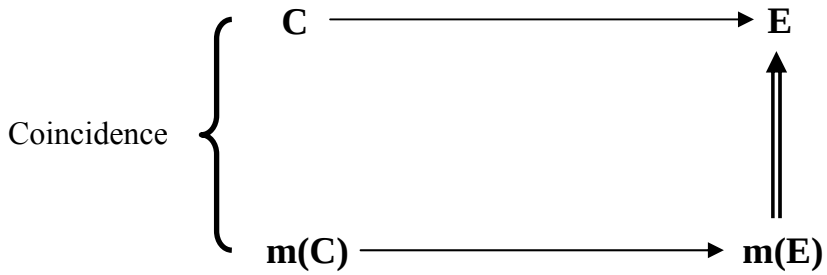
effects. If it is not the scientist's decision to intervene in the quantum world that causes Pow!, then neither is it my decision to type this sentence that causes my fingers to move. And that is just *false*.

For those who do not wish to rely on transitivity, or indeed Witmer's points about intervention, there is another, to my knowledge universally overlooked, reason why transmission principles are not needed in the causal argument. Sufficiency, you will recall, is needed to render the co-occurrence of mental and physical causes non-coincidental. What is overlooked is that nothing in the way we think about coincidences entails that such co-occurrences are unacceptably spooky only if the two causes are causes *of the same effect*. Suppose, for the sake of argument, that m(E) above is sufficient for E. Given $C_P$, we know that m(E) has a physical cause. Given $E_M$, we know that C causes E. But that is all we need, for we now know this much: that whenever a mental event causes a behaviour, there is a physical event such that it causes those physical events that are sufficient for the behaviour. Now this strikes me as just the sort of coincidence that we need sufficiency to rule out. Notice also that the relationship between m(E) and E does not have to be as strong as sufficiency. Again for the sake of argument, let m(E) partially compose E. Now we are faced with the coincidental prospect that whenever a mental events causes a behaviour, there just happens to be a physical event that causes *part* of that behaviour. And the question, as before, is this: how are we to explain the co-occurrence of the mental and physical events without appealing to a sufficiency relationship between them?[138]

It is tempting to suppose that we can go still further, and generate the required coincidences without *any* mention of the effects of the mental and physical causes. We might appeal to the simple fact that we have a mental and a physical cause occurring at the same time with a greater probability than if they were *un*related. But it is not clear to me that this suggestion can work. The reason is that there is no *particular* physical event that has to occur at the same time as, say, my decision to

[138] See section 2.3 for discussion of this issue.

make some tea. But then we are faced with the prospect of assigning a probability to the joint occurrence of my decision and...what? Some physical event or other? The only thing physical events that occur at the same time as mental events of this type have in common is that they all cause events that are synchronically sufficient for tea-making behaviour. Our coincidence consists not merely in the co-occurrence of a mental and *some* physical event, but in the fact that they are co-occurring causes of two synchronically related events. We can depict the coincidence like this:




Another way of putting the point is this: the failure of causal transmission principles (if they are indeed false) means that m(C) does not cause E. But it does not mean that m(C) and E are *unrelated*. Indeed, I think it clear that they are quite intimately related; m(C), as cause of m(E), is nomically related to E given the sufficiency of m(E) for E. It is a matter for us to choose what to call the relationship between m(C) and E – if we endorse counterfactual theories of causation, then we must find another name for it. Sturgeon calls it 'inducing'; I can't think of a better name, so let's say that m(C) induces E. Define 'sufficient induction' as follows:

*x* sufficiently induces *z* iff: (i) *x* is causally sufficient for *y*;
(ii) *y* is synchronically sufficient for *z*.

Now it is surely just the kind of coincidence that intuition cannot tolerate that, whenever I *cause* my arm to move by deciding to move it, there is a physical event simultaneous with my decision that *induces* the very same movement. Unless, of course, the physical inducer is sufficient for the mental cause. This suggests an obvious reformulation of $O_D$ to accommodate coincidental inducers as well as causes:

$O_D^3$: if an event *y* has a sufficient cause or inducer *x* at *t*, then no event *x*' is also a cause or inducer of *y* at *t* unless (i) *x* is sufficient for *x*', or (ii) *x*' is sufficient for *x*.

The central conclusions of this chapter are as follows: (i) an argument that relies on causal transmission principles is open to easy rebuttal by those who endorse any one of a number of theories of causation, so it is better not to rely on such principles; and (ii) the causal argument does not have to rely on such principles, and so can retain its neutrality as to the choice of a theory of causation. In the next chapter we will see that the familiar 'causal exclusion argument' relies on a particular conception of the causation. As such, it is bound to be weaker than the causal argument, as a defence of the exclusion argument will depend on a defence of the relevant theory. I will argue, however, that things are much worse than that for the exclusion argument, as in fact the theory of causation upon which it relies is demonstrably false.

# 5. Against the Causal Exclusion Argument

This argument, if successful, will show that only the identity of mental and physical events enables us to make sense of mental causation. It is a highly controversial argument, not least because (on reasonable assumption) it entails a metaphysic of mind – type identity – that many regard as sufficiently problematic to license a *reductio* of the argument. In particular, as we have already seen in 2.3, the identification of mental and physical property-instances, when combined with multiple realization, entails eliminativism. I will not, however, rely on the *reductio* that might be so constructed in order to defeat the exclusion argument. My own counter argument will instead be directed at the theory of causation that I take to be essential to the exclusion argument. That the exclusion argument does depend on a theory of causation is hinted at by certain authors, though not, to my knowledge, treated with any great rigour. As we saw in chapter 4, it is a virtue of the causal argument that it need not rely on any particular theory of causation. My argument in this chapter will be that not only does the exclusion argument lack this virtue, it also has the added vice of relying on a theory of causation that is just plain *wrong*.

I will argue as follows. First, I will examine Kim's 'supervenience argument', and through this introduce the role of the concept of 'causal work' in exclusionary reasoning. In 5.2 I argue that given the theoretical work those who appeal to the concept expect it to do, the most plausible understanding of causal work is as physical work of some kind. I give an account of causation in terms of physical work by formulating general principles based on the properties of physical work. I show how a stronger form of the no-overdetermination premise (which Kim relies on in the supervenience argument discussed in 5.1) can be derived from these principles. Correspondingly, I formulate the causal exclusion argument as a version of the causal argument based on this stronger no-overdetermination rule. In 5.3, I give a general account of the causal exclusion problem in terms of the theory of causation outlined in 5.2, and give a brief taxonomy in 5.4 of possible responses to the problem so formulated. I show that several authors respond to the problem by implicitly rejecting

the causal work theory of 5.2. In 5.5, I argue that there are clear counterexamples to the principle they reject, and argue that since the problem of causal exclusion (and the corresponding causal exclusion argument) relies on this principle, there is no problem of causal exclusion.

## 5.1. The supervenience argument

Let's begin by taking a closer look at Kim's argument against supervenient causation, which I introduced in 3.3. In a series of papers, Kim argues that supervenience, the initial introduction of which can be seen as an attempt to bring the mind into the causal structure of a physical world, in fact renders mental causation utterly mysterious. This poses a dilemma, which Kim states in the following way:

> If mind-body supervenience fails, mental causation is unintelligible; if it holds, mental causation is again unintelligible. Hence mental causation is unintelligible.[139]

Failure of supervenience means that the mental isn't 'anchored' to the physical in any way that would make intelligible the fact that mental events have physical effects. Understood in this way, the first horn of the dilemma is not importantly different to the coincidence worries that motivate $O_D$ in our causal argument of 3.4. We need not concern ourselves with arguing this point, however – even if this is not what Kim has in mind, we know that if physical events are *not* sufficient for mental events, then the world is full of coincidences, and this in itself is unintelligible. We concern ourselves with the second horn of Kim's dilemma. Why does Kim think that if supervenience *holds*, then mental causation is unintelligible? In what follows, I will refer sometimes to supervenience relations between token events, sometimes to sufficiency relations. For my part, as I argued in 1.4, I think that non-causal relationships between token events are not best characterised in terms of supervenience. When I speak, somewhat loosely, of the 'supervenience base property' of a particular property instantiation, I

[139] Kim [1998] p.46.

do so only for consistency with Kim's terminology, and do not intend by such talk to backtrack on any of what I said in 1.4.

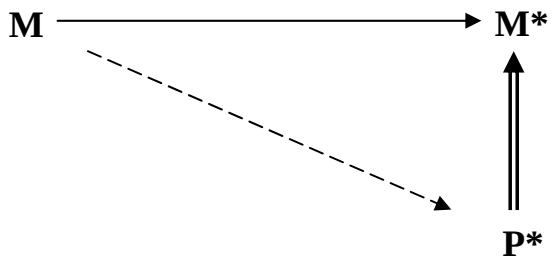
Kim presents us with two distinct arguments for the conclusion that supervenient properties are causally inefficacious; each is a two-stage argument, and they have their first part in common.[140] The first part in each case is in essence the 'downwards transmission argument' we saw Witmer running in 4.2. The purpose of this part is to show that same-level causation presupposes what Kim calls 'downwards causation'. There is no causing the instantiation of, say, a supervenient property, without causing the instantiation of its base property. The second parts of each argument are designed to show that downwards causation gives rise to an unacceptable causal competition between mental and physical properties, and to show in addition that this is a competition that the mental property must lose. Each second half relies on the completeness of physics, and a 'causal exclusion principle'. The first of the two second halves is a somewhat curious argument, which I term the 'upwards-downwards transmission argument'. Its curiosity consists in the fact that it is unnecessarily circuitous on the assumption that $C_P$ is true; I include it here as Kim sometimes runs it *without* assuming $C_P$, to show that emergentism is untenable. We return to this version of the upwards-downwards argument in 7.3. The other second half is a much simpler, and far more powerful argument. Although both of Kim's second halves could properly be termed 'causal exclusion arguments', for expository reasons I reserve this term for the second of them.

Part 1: The downwards transmission argument

Kim actually gives two arguments for the conclusion that same-level (supervenient) causation presupposes downwards causation, but I think that they can both be properly termed 'downwards transmission arguments'. Only the second, however, is equivalent to what we termed the downwards transmission argument in 4.2; I present both here for completeness. *Firstly*, Kim feels that there is a tension between the

[140] The argument occurs, in various forms, in his [1992b], [1993b], [1998] pp.40-7, [1999a] pp.32-4, and in its clearest form in his [2003] pp.155-9. I will rely on this latter in my exegesis.

causation of one supervenient property-instance M* by another such property-instance M, and the synchronic sufficiency of M*'s base property P* for M*. We seem, in this case, to have determination from two directions: M* is *causally* determined by M, and *non*-causally determined by P*. Kim argues that the only way to make sense of this 'double-determination' is to suppose that M causes M* by causing its base property. We can use the diagram below to depict what is going on.




Kim argues that there is no way to reconcile the non-causal determination of M* by P* with the causal determination of M* by M, unless we suppose that M causes M* by causing P*, hence the downward arrow. Now as I have already argued, causal transmission is dubious. Why does Kim think it is the only way to explain how M* can have distinct diachronic and synchronic determinants, M and P* respectively? Frankly, I'm not sure, although I think it may be due to the fact that Kim does not at this stage appeal to any causal relationship between *M's supervenience base*, P, and P*. If M either has no supervenience base, or has a physical base property P which does *not* cause P*, then I agree it would *very strange indeed* that M managed to cause M*, which depends on P*, without also causing P*; after all, *something* has to cause P* in order for M* to be instantiated. The situation depicted above would look decidedly odd without the dotted downwards causation arrow. However, suppose we draw in P, and accept that P causes P*; then there is nothing obviously *mysterious* in the thought that M causes M* *without* causing P*. For instance we might maintain, as Yablo does, that causation is an intra-, not an inter-level relation. On this account, P causes P*, and M causes M*, but there is no diagonal causation. Why does Kim not appeal to P as cause of P*, and resolve the tension that way? I must confess I don't know, although I think it may just be that Kim is looking for a general argument form

that tells equally against supervenience physicalism *and emergentism*. And as we shall see in chapter 6, there are forms of emergentism according to which $C_P$ is false, so that no arrow can be drawn from P to P*.[141] Kim goes on to give a second argument to the effect that M must cause P*. Whereas the first argument claims that downwards causation is the only way to *understand* simultaneous causal and non-causal determination; the second claims that *the only way to cause* the instantiation of a supervenient property like M* is to cause its base property. He says this:

> To relieve a headache, you take aspirin: that is, you causally intervene in the brain process on which the headache supervenes. That's the only way we can do anything about our headaches.[142]

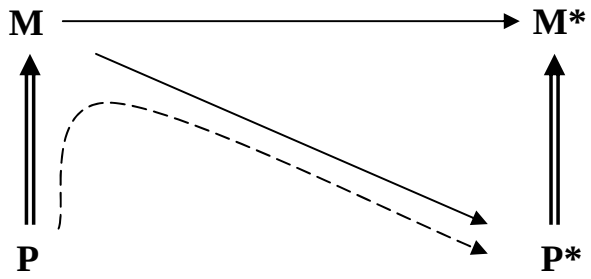
Now this has nothing to do with resolving any purported tension between the synchronic and diachronic determination of M*; indeed, Kim intends this argument to appeal to those who don't *see* the tension. And this, of course, is just the same argument that we saw Witmer running in 4.2 to generate causal competition as a reply to Sturgeon's charge of equivocation. Now it should be clear that I do not think reliance causal on transmission is a virtue. Still, let us grant for the sake of argument that if M causes M*, then M causes P*. This concludes the first part of the supervenience argument; each of the alternative second parts is designed to show that M's supervenience base P *also* causes P*, and to argue that as a result, given $C_P$, P displaces M as the true cause of P*.

Part 2(a): The upwards-downwards transmission argument

This is what Kim [2003] terms 'Completion 1'. The argument to follow also relies on causal transmission, this time in order to establish that P causes P*. Once this much established, Kim proceeds to argue that P *excludes* M as cause of P*. The argument to this conclusion rests on the assumption that M supervenes on P, and holds (not implausibly) that this entails that M≠P; a cause cannot exclude itself as cause of its effect. The argument also rests on $C_P$, and a 'principle of exclusion', which I will state

[141] We will examine the application of the argument to these forms of emergentism in 7.3.
[142] [1998] pp.42-3.

below. Kim argues that since P is sufficient for M as its supervenience base, and M causes P*, it follows that P causes P*. The argument can be depicted like this:




Now Kim seems to think that P's synchronic sufficiency for M *qua* M's supervenience base, plus M's causal sufficiency for P* as established in part (1), is sufficient for P's causal sufficiency for P*. If this is true, then we can draw in the curved dotted arrow. Kim seems to have two primary reasons for thinking this arrow can be drawn in.[143] *First*, Kim holds that if nomological sufficiency is sufficient for causation, then since P is sufficient for M, and M is sufficient for P*, by transitivity of sufficiency it follows that P is sufficient for P*, and so qualifies as its cause. I grant Kim that *if* causation is understood in these terms, then P qualifies as P*'s cause. Whether or not causation should be so understood is, however, a highly contentious matter. *Second*, Kim holds that if counterfactual dependency is sufficient for causation, then since (Kim claims) the closest possible worlds in which P does not occur will be worlds at which M does not occur, then given that P* counterfactually depends on M, P* counterfactually depends on P. It is unclear to me that Kim is right that the closest ¬P worlds are ¬M worlds; I argued in 4.3 that if P is relatively fragile, then a world where an alternative realizer P' of M will be closer to actuality than a world at which no realizer of M occurs. No matter, I will grant Kim that P* counterfactually depends on P, for the sake of argument. From this it follows that *if* counterfactual dependency is sufficient for causation, then P qualifies as cause of P*.

[143] See Kim [1998] pp43-5 for the most complete statement of his reasons (at least that I am aware of) of the argument described here. I will not address these arguments in the present chapter. I will return to a version of it, directed against the cogency of emergentism, in 7.3.

Once again, whether causation is to be understood in these terms is highly contentious.

Now the above argument is a highly circuitous route to the conclusion that P causes P* – in effect, Kim thinks that P must cause P* because it (P) is non-causally sufficient for something (M) that causes something (M*) for which P* is non-causally sufficient. Talk about pushing causation around! Still, let us grant Kim that on the assumption that M≠P, there is a 'causal competition' between M and P as cause of P*. And it is this 'causal competition' that matters to Kim: we have two putative causes (P and M) of the same effect P*. The next stage of the supervenience argument is to appeal to a version of the causal exclusion principle, and apply it to our two competing causes. Here is the version to which Kim appeals:

> $E_X$ No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.[144]

Given this principle, we can conclude that at most one of M and P is a cause of P*. But which one are we to choose? This is where the completeness of physics comes in.[145] We can neither treat M and P as jointly causally sufficient for P*, nor treat M as the cause of P* and deny that P causes it. In the former case, we are faced with a physical event, P*, for which there is no sufficient physical cause – it is P and M *together* that are sufficient, and this violates $C_P$. In the latter case, either P* has no physical cause at all – which clearly violates $C_P$ – or else there is some P' that causes P*, in which case the same problematic causal competition obtains between M and P'.

---

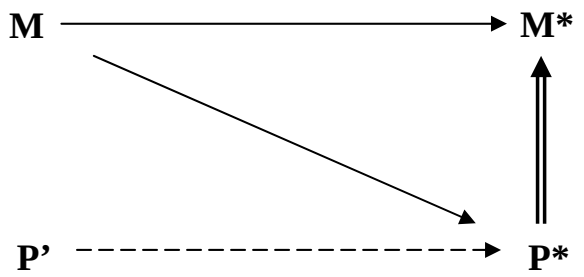
[144] Kim [2003] p.157. This principle is clearly much stronger than the exclusion principle we discussed in 3.3; I comment further on this below. In passing, I note the following *ad hominem*: $E_X$ does not fit well with the premise, which Kim appeals to in the upwards-downwards argument, that either nomic sufficiency or counterfactual dependency are sufficient for causation. In fact, such accounts of causation are normally appealed to by 'causal compatibilists' seeking to show that principles such as $E_X$ are false, and that events can have more than one cause! I return to the connection between compatibilism and theories of causation in section 5.4, when I discuss possible solutions to the exclusion problem.

[145] See Kim [1998] p.44-5, and [2003] p.158 for explicit appeals to $C_P$ in the upwards-downwards transmission argument.

We can now conclude this version of the supervenience argument as follows: P excludes M as cause of P*. We are forced, given $C_P$, to choose between either (i) retaining M as a property distinct from P and giving up on $E_M$; or (ii) maintaining $E_M$ and identifying M and P. This is the *causal exclusion problem*. The choice seems to be epiphenomenalism or identity, and neither is appealing. Epiphenomenalism flies in the face of all the *prima facie* evidence for $E_M$ suggested in 3.1. Given the overwhelming plausibility of multiple realization, identity leads to eliminativism, as we saw in 2.3. I find it strange that Kim appeals to $C_P$ to justify choosing P rather than M as the cause of P* in the present form of the supervenience argument – for if $C_P$ is true, then we do not need the (arguably) somewhat spurious upwards-downwards transmission argument in order to conclude that P* has a physical cause. Nor, as we shall see, do we need to rely on the supervenience of M on P.

Part 2(b): The causal exclusion argument

This argument, like 2(a), depends on the downwards transmission argument, but appeals neither to the supervenience of M on P nor the contentious claim that either counterfactual dependency or nomological sufficiency is sufficient for causation. We can depict this argument as follows:




In this argument, we appeal to $C_P$ to show that P* has a sufficient physical cause P'. In the above diagram, P' must occur at the same time as M. This is because $C_P$ tells us that any physical event that has a sufficient cause at *t*, has a complete sufficient physical cause *at t*. Since the downwards transmission argument tells us that P* has a cause occurring at the same time as M (viz. *M*), we can infer only that it also has a complete sufficient *physical* cause occurring *at the same time* as M. Nothing in this

formulation of the argument requires as a premise that P' is M's supervenience base. It plausibly will be, on the assumption that M *has* a physical supervenience base; but the nice thing about this version of the argument is that we do not have to worry about whether this is so. Now once again, given that M≠P, we have causal competition for P*. Once again we can just appeal to $C_P$ to conclude that, given $E_X$, M can not be a cause of P*. To conclude otherwise would be to violate $C_P$. And if M is not a cause of P*, then neither, by the argument of part (1), M a cause of M*. The only option open to us if we want to retain $C_P$, $E_M$ and $E_X$ is to identify M and P. This concludes the supervenience argument.

An obvious response at this point is, why believe $E_X$? Notice first how much *stronger* $E_X$ is than the principle of causal exclusion we discussed in 3.3. The principle we discussed there made explicit reference to the *independence* of the two causes, and as we saw, this fact exempted the principle from application to causes standing in dependency relations such as supervenience or sufficiency. There are plenty of dependency relations other than identity. The principle invoked here has no mention of the independence of the causes – rather, $E_X$ claims that if both A and B are sufficient causes of C, then except for (rare) cases of overdetermination, A=B. We can of course agree that there is no way we can treat all cases of mental causation as genuine overdetermination, for that is to invoke the possibility of widespread overdetermination. However, it is unclear why we should endorse the stronger principle that denies that any numerically non-identical events (dependent or not) can have the same effect. For that an event has two causes, one of which *depends* on the other, is clearly not overdetermination of the problematic kind. But now recall the three overdetermination worries we briefly discussed in 3.3, of which the coincidence worry was just one. Recall that Kim thinks that there are the following problems with the view that P* might have two causes: (i) that P* is overdetermined, which is absurd; (ii) that M is redundant as a cause of P*, and so can be dispensed with, and (iii) that there is no causal work left for M to do, given P*.

The first of Kim's problems, we can dismiss: there is nothing *absurd* in the supposition that P and M both cause P*, for *given* that P and M both cause P*, we can infer by $O_D^3$ that P is sufficient for M, hence that this is not a case of genuine overdetermination. Sufficiency, as I have argued, is all that is required to render the co-occurrence of M and P non-coincidental. The second of Kim's problems we return to in 6.2, where I will argue that only if we endorse a very weak criterion for redundancy (in the form of a very strong criterion for *novelty*) does M count as redundant. And the third problem, we will now examine in detail. First, it may be asked why I do not simply deny the problem, as relying on the kind of transmission principles whose truth I spent most of chapter 4 questioning. If I thought that the downwards transmission argument discussed above were the only route to the conclusion that there is no causal work for mental events to do, then I might well reject it on those very grounds. However, as I will show in the next section, we do not need to assume *any* transmission principles in order to leave the mental shorn of any "causal work". In what follows, I will show that given certain plausible further assumptions about the nature of *causal work*, supervenient properties do not do any of it. I will not rely on any form of causal transmission argument. In 5.2 I will attempt to outline some general truths about causal work, and formulate them as principles. I formulate these principles based (*inter alia*) on certain things that Kim (and others) have said. I take it to count in support of my having got Kim right on causal work that the principles I formulate will enable us to *derive* an exclusion principle equivalent to $E_X$. Through this, I will proceed in 5.3 to show how these principles, when combined with $C_P$ and $E_M$, give rise to a highly general causal exclusion problem.

## 5.2. Some principles of 'causal work'

What is causal work? I will not attempt to give a definition; instead, I will describe roughly the sort of thing it is supposed to be. I will then briefly summarise several extant theories according to which, in broad outline, causation involves doing some causal work. Finally, I will formulate what I take to be three central theses describing causal work, as a preamble to the causal exclusion problem of 5.3 (referred to below as $P_L$, CW, and $T_{CW}$). Suppose first that a certain amount of causal work is required in

order for an event to happen. If you want to build a wall, then you need to move bricks about, mix cement, and so on. In general, let's say, if you want to make it so that a substance has a property at a time (which is exactly what you do when you build a wall), then you will have your work to do. I will refer to this work, in what follows, as the *causal work required* for the occurrence of an event. Physicists understand 'work' in terms of energy transfer. In particular, if a force is required to move a body, then the action of the force transfers energy to it, and the energy transferred is equal to the work done.[146] This accords well with a certain intuitive way of thinking about causation – if you want to move something around, or restructure the configuration of a group of things, then you will need to supply some energy. Intuitively, we might reason as follows: causation involves making changes to the world; but change takes work, and work is the transfer of energy. So causation involves the transfer of energy. Understanding causal work as *physical* work has the advantage of giving us a ready-made account of an otherwise ill-understood notion. However, caution is necessary: in particular, the view that causal work is physical work does *not* entail that doing the causal work required for an effect is either necessary or sufficient for being its *cause*. This is an important point, as it enables us to maintain the (in my view plausible) connection between causal and physical work, without endorsing the view that causes must do physical work on their effects. I will argue in 5.5 that doing the causal work required for the occurrence of an event is not a necessary condition for causing it. For the moment, we turn to accounts of causation that *do* hold that it can be reduced to work. Through this, we will derive principles that enable us to set up the causal exclusion problem in 5.3.

The intuitive conception of causation as physical work finds voice in Fair, who maintains that the relata of causation are objects, and that causation can be identified with a flow of energy or momentum from cause to effect.[147] Kistler, on the other

[146] Specifically, physicists define the work done on a body along a path as the integral of the scalar product of the force with the infinitessimal of the distance through which the force acts. What is important about this definition for my purpose is that physicists define work in terms of energy transfer.
[147] Fair [1979].

hand, takes Fair's reliance on specific quantities to be arbitrary (and problematic on other grounds, which I will not go into here), preferring to account for causation in terms of the transfer of whatever quantities obey physical conservation laws.[148] Thus Kistler defends the thesis that

> "[t]wo events *c* and *e* are connected by a causal relation if and only if there exists a conserved quantity Q which is exemplified by both *c* and *e* and of which a particular amount A is transmitted between *c* and *e*."[149]

Dowe has developed a similar account, but takes the notion of causal *process* to be prior to that of causation. Dowe construes a causal process as the worldline of an object that *possesses* a conserved quantity, and a causal *interaction* as an intersection of world lines involving the *exchange* of such a quantity.[150] Put simply, a causal interaction occurs at the intersection of two (or more) causal processes, in which the processes concerned undergo changes in the values of whatever quantities. Current science is our best guide to which quantities are conserved, so there is good reason to think that causal interactions involve exchanges of mass-energy, charge, and so on. I will not attempt to determine which of the above theories is the most plausible. Instead, I will follow Kistler in assuming, for present purposes, that 'causal work' is the transfer of some conserved quantity from cause to effect. It should be clear that the central arguments of the remainder of this chapter will go through *mutatis mutandis* on any of the alternative accounts mentioned.

Clearly, all the theories described above construe causation, broadly, in terms of a *process* that connects cause and effect – correspondingly, they (and other variants) are commonly known as *process accounts* of causation. The important point is that such accounts all maintain that there is an intrinsic, physical connection of some kind between causes and effects.[151] There is clear evidence in Kim's work that he thinks of

[148] Kistler [1998].
[149] Kistler [2001] p.115.
[150] Dowe [2000].
[151] An 'intrinsic relation' can be understood as a relation holding between the members of an n-tuple solely in virtue of the intrinsic properties of the objects involved, construed according to the account of 'intrinsic' I endorsed in 1.1. Counterfactual dependency, by way of contrast, can be made to hold or fail

causation in a similar way to the authors mentioned above. For instance, in one discussion of the supervenience argument, Kim says this:

> [T]he observed regularities between M-instances and M*-instances….are by no means accidental….However, if we understand the difference between genuine, productive and generative causal processes…and the noncausal regularities…that are parasitic on real processes, then we are in a position to understand [that] [i]n the case of supposed M-M* causation, the situation is rather like a series of shadows cast by a moving car: there is no causal connection between the shadow of the car at one instant and its shadow an instant later….[152]

This passage clearly indicates Kim's reliance on a process account of causation, as he explicitly equates genuine causation with 'production' and 'generation'. But why, on such an account, does the relationship between M and M* turn out to be a pseudo-process? In what follows, we will see how, given a few reasonable assumptions, Kim's position is virtually mandatory given a process account of causation. In the remainder of this section, I will outline general constraints on causal work, and formulate principles accordingly.

Effects, on a process account, are literally generated, or produced by their causes via transference of some sort, which we will take in what follows to be transference of a conserved quantity from cause to effect.[153] Problematically, such transfer will not necessarily occur where a cause has its effect via transitivity; *x* might transfer a quantity to *y* and *y* to *z*, without the transfer of any particular quantity at all from *x* to

---

by adding or removing objects to the world of a pair of events related as cause and effect – you can add in a backup or 'redundant' cause such that it is only active if the actual cause fails to have its effect. But then despite no change in the intrinsic properties of cause or effect, the effect will fail to counterfactually depend on the cause. In general, causation construed as counterfactual dependency will not be an intrinsic relation.

[152] Kim [1998] p.45. In fn.28, Kim makes clear he endorses Salmon's distinction between causal processes and pseudo-processes. Salmon maintains that only a causal process is capable of transmitting a mark. Mark transmission is analyzed in terms of the preservation, without intervention, between A and B, of changes in the characteristics of the process due to a local interaction at point A. See Salmon [1984] pp.450-2.

[153] Talk of cause transferring a quantity *to its effect* is, problematic and ought not to be interpreted literally. Such talk would seem to suggest that the effect already exists, in which case that something else *already* caused it. Transference "to the effect" is better understood to take place between property-instances, where the first of which gives up a certain quantity of a physical property, and the second the receives the property, *becoming* the effect in question in virtue of the property gained.

*z*. You make a phone call asking someone to build you a shed, thereby transferring a quantity to them down the phone line, and they proceed to transfer quantities to the pile of wood in your garden. We know *a priori* (relative to empirical investigation of the nature of the underlying process, and assuming transitivity) that your phone call causes the shed to be built; but it is not likely (although I suppose it is possible) that any amount of any *particular* quantity survives the journey down the phone line and into the shed. In order to explain transitivity, then, we must distinguish two types of cause. Following Kistler [2001] let us say that 'direct' causes produce or generate their effects *proximally* by transfer of a conserved quantity; while 'indirect' causes are related to their effects via chains of direct cause-effect pairs. The causal work required for the occurrence of a given effect, on the present understanding of 'causal work', will only be done by its *direct* cause. Our first principle concerns direct causation, and reflects the claim that proximal causes generate, or produce their effects. Correspondingly, we may refer to it as the 'generativity of direct causation', and define it thus:

$G_C$ An event *x* is a direct sufficient cause of an event *y* iff *x* does all the causal work required for the occurrence of *y*.

Direct sufficient causes, according to this principle, have their effects by transferring certain conserved quantities, and so doing the causal work required to bring about the effect. We will also need a weaker principle, for full generality, to cover cases of *in*sufficient direct causation. We can formulate the weaker principle like this:

$G_C$' An event *x* is a direct cause of *y* iff *x* does some of the causal work required for the occurrence of *y*.

We can now proceed to give a principle of both sufficient and insufficient causation in general, based on our two generativity principles. This is the business of our third principle. Define a *causal process* as a temporally ordered sequence of events each of which is a direct cause (as defined in $G_C$') of the next. A natural progression is to

define a *sufficient causal process* as a causal process in which each event is a *direct sufficient cause* (as defined in $G_C$) of the next. For instance, you throw a brick through a window, your throw is a sufficient cause of the brick's motion, the motion is sufficient for the impact, and the impact is sufficient for the shattering. It is this chain of sufficient causes that makes your throw a sufficient cause of the shattering, rather than a partial or contributory cause. We can now give the following necessary and sufficient condition for causation, covering both the sufficient and insufficient varieties. Following Schaffer [2001] let's call this the *process-linkage* theory:[154]

$P_L$ An event *x* is a (sufficient) cause of an event *y* iff *x* and *y* are parts of a (sufficient) causal process in which *x* occurs prior to *y*.

This principle tells us that *x* sufficiently causes *y* just in case *x* and *y* are part of a temporally ordered series of events, each one of which does all of the causal work required for the next. For insufficient causation, simply delete the parenthesised occurrences of 'sufficient'. Direct causation will be a limiting case of causation as defined by $P_L$, where *x* and *y* are adjacent events in the series. A further constraint on causal work is evident: it is not the sort of thing that gets done *twice*. If the causal work required involves the exchange of a conserved quantity, for example, then once this quantity has been exchanged, it doesn't get re-exchanged. Any more quantities that are exchanged will be parts of different causal processes. Put more simply, once you have done the causal work required to build a wall, there just isn't anything left to do. Two people can certainly build the same wall, but they do not replicate causal work, they *share* it. Let us call this view the 'causal work principle', CW, and define it as follows:

CW The causal work required for a given effect is done at most once.

[154] I borrow Schaffer's terminology only; I do not intend by this to imply that Schaffer would accept the version of the process-linkage theory I am describing here.

The final thesis about causal work I would like to endorse concerns how causal work is related to *sufficiency* as detailed in 1.4. Consider again the process of building a wall. You partake in this process by cementing individual bricks in place, causing a specific aggregate configuration of bricks to exist. The aggregate is plausibly not identical to the wall, due to their (plausibly) different modal properties; it is, however, *sufficient* for it, and the wall is clearly "nothing over and above" the aggregate. Once you have finished assembling the aggregate, then, your *causal work* is done. It isn't as though you assemble the bricks, take a couple of days off, then come back to finish the job. What goes for walls goes for synchronic sufficiency generally.[155] The lack of any more causal work to be done, at least on the present understanding of causation, is *necessary* if the sufficiency is to *be* noncausal. Call this view 'the transmission of causal work', and define it thus:

$T_{CW}$ If $x$ is a direct sufficient cause of $y$ and $y$ is synchronically sufficient for $z$, then $x$ does all the causal work required for the occurrence of $z$.

Now from these principles of causal work, we can make some interesting derivations. First, note that we can derive the upwards causal transmission principle of 4.1 from $G_C$ and $T_{CW}$. According to the first principle, doing the causal work required for an effect is a sufficient condition for directly causing it. But now from $T_{CW}$ it follows that the direct cause of $y$ directly causes $z$. Thinking of causation in terms of pushing, pulling and 'oomph' makes it wholly unmysterious why Kim is so ready and willing to 'push causation around'. I will not take the time to do so, but I am sure a similar principle could be formulated to enable us to derive the downwards transmission principle as well. After all, the causal work required for $y$ and $z$ is the same. If causation is thought of in terms of process and transference, then it seems that causal transmission principles are virtually unavoidable. Still, I will not assume them in anything to follow, for I have no need of such principles. More importantly than the derivation of transmission principles, however, is that we can derive from our causal

[155] At least the noncausal variety. I briefly discuss the complications posed for $T_{CW}$ by simultaneous causation when I consider rejection of $T_{CW}$ as a solution to the exclusion problem in 5.4.

work principles the very strong *principle of causal exclusion* often (at least tacitly) appealed to in formulations of the causal exclusion problem. We can derive this principle from $P_L$, CW, and $T_{CW}$; here is how the derivation goes.

Suppose that *y* has a sufficient cause *x* at *t*. Then by $P_L$, there is a sufficient causal process linking *x* and *y*. From this, it follows by definition of 'sufficient causal process' as a chain of direct sufficient causes, and the definition of direct sufficient causation in $G_C$, that *x* does all the causal work required for the event $x^+$ that follows it in the series. Now consider some event M that also occurs at *t* such that $x \neq M$, and let us ask whether or not M can also be a cause of *y* at *t*. If it is, then by $P_L$ we know that there must be a causal process linking M and *y*. By CW we know that the causal work required for $x^+$ gets done one time only, and all of it gets done by *x*. And the same follows, *mutatis mutandis*, for the causal work required for $x^{++}$, and so on until we reach *y*. But from this it follows that there is no causal process *whatever* (sufficient or otherwise) linking M to any of the events in the process leading from *x* to *y* – for if there were, then *at least some of* the causal work required for one of these events would be done twice. And so by $P_L$, since there is no causal process linking M to *y*, M is not a cause of *y*. Might M be a cause of some event for which one of the events in the process from *x* to *y* is noncausally sufficient? No. By $T_{CW}$, all this causal work is also taken care of by the events in lower-level process. Notice that the inefficacy of M follows, given the process account, on the assumption of its non-identity with *x*. It follows that $P_L$, CW and $T_{CW}$ entail the following variant of our principle $O_D{}^3$ of 4.4, got by replacing sufficiency with identity:

> $O_D{}^4$: if an event *y* has a sufficient cause or inducer *x* at *t*, then no event *x*' is also a cause or inducer of *y* at *t* unless $x=x'$.

Notice that this derivation depends only on $P_L$ as a *necessary* condition for causation – no inference drawn here requires the *sufficiency* conditional 'if events are linked by a process, then they are causally related.' Rather it is the other way around: the appeal to $P_L$ is of the form, *if* A causes B *then* there is a causal process linking A to B. That

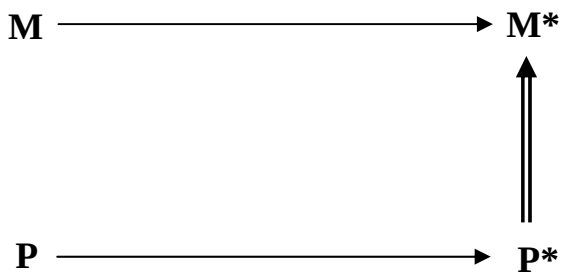
$P_L$ could be a necessary condition for causation is precisely what we will deny in 5.5, for $P_L$'s putative necessity, as we shall see, raises what I consider to be an insurmountable stack of counterexamples to the process-linkage view.

$O_D^4$ is, I think, much more plausibly termed an 'exclusion principle' than the version due to Kim, which we examined in 3.3. Given that *x* causes *y*, $O_D^4$ literally *excludes* any *other* putative causes of *y*. And $O_D^4$ is equivalent to $E_X$, except that the former covers sufficient causal *inducement* as well as sufficient causation. A troublesome tendency in the literature on mental causation has been to treat principles such as $O_D^4$ as if they were of a piece with coincidence-motivated principles such as $O_D^3$. Kim, for instance, in the passage I quoted in 3.3, equivocates between talking of a lack of causal work for M to do, and the absurdity of causal overdetermination. As I take it I have shown, however, the principles are anything but equivalent. Combined with $E_M$ and $C_P$, $O_D^4$ entails that mental and physical events are *identical*, and so on the present conception of events entails that mental and physical *properties* are identical.[156] Now as we saw in 2.3, type identity and multiple realization do not combine happily. As such, in the remainder of this chapter, we will look at ways of resisting $O_D^4$. The tendency to regard $O_D^4$ as really just another way of affirming the absurdity of causal overdetermination is a troublesome one not least because it encourages the view that anyone who *denies* $O_D^4$ is endorsing an absurd position! Since $O_D^4$ follows from the causal work principles given above, then denying it will clearly involve a denial of at least one of these principles. For this reason, in the interests of clarity, I will not formulate the exclusion problem in terms of $O_D^4$. Instead, I will formulate it in terms of causal work principles, and so facilitate a more complete taxonomy of responses to the problem.

[156] It should be noted that we do not *require* Kim-events in order to deduce type identity from this new causal argument. We could run the argument with Davidson-events, for instance; but in that case, we would have to run it twice, once for events, then again for the properties in virtue of which they are related as cause and effect. An advantage of Kim-events is that we only need to run the argument once – if the events are identical, so then are their constitutive properties. See 2.3 for further discussion of this point.

## 5.3. The causal exclusion problem

As I said in 5.1, Kim's refusal to assume $C_P$ results in a lot of unnecessary 'pushing around' of causation. In order to avoid this, we will rely on $C_P$ to generate the problem. Let us once more consider a putative case of mental causation M-to-M*, where there is a physical event P* that is noncausally sufficient for (say) a behaviour M*.[157] Suppose that M and P occur at the same time. Unlike Kim, we will not assume that P is sufficient for M. Instead, we will appeal to our principles to show that if M causes M*, then M=P. Kim's argument assumes that M supervenes on P in order to generate a problem for supervenience; we will show that the only relationship between M and P that allows M any causal work to do is identity, which raises exactly the same problem for supervenience, *a fortiori*. Consider the causal diagram below:




By $E_M$, we know that M causes M*. Now by $C_P$, (assuming P* to have a sufficient cause) we know that P* has a sufficient physical cause, P. But then from $P_L$ it follows that a sufficient causal process connects P with P*. Applying $P_L$ to M's causing M*, we know that there must be a causal process connecting M with M*. Again, by definition it follows that M does at least some of the causal work required for its proximal effect $M^+$, $M^+$ does the causal work required for $M^{++}$, and so on forwards until we reach M*. By $T_{CW}$, however, we know that all the causal work required for M* is done by the event that directly causes P*, let's call it P*-. But all the causal

[157] As we saw in 5.1, Kim sets up the situation in terms of 'mental-to-mental' causation, and it is unlikely that he would count the causation of behaviour as an instance of this. Nothing turns on this difference, for all the central arguments go through *mutatis mutandis* regardless of how M* is conceived. I appeal to behaviour here because I do not wish to assume the supervenience of mental properties on physical properties. However, since we already have good functional reductions of those behaviours that involve bodily movements to physiology, I take to be fairly uncontroversial that the assumption of *behavioural* supervenience is less problematic. See chapter 2 for details for the relationship between functional reduction and supervenience.

work required for P*- is done by P*--, and so on backwards until we reach P. But it now follows from CW that unless M is identical to one of the events in the process from P to P*, M cannot be a cause of M*. For M to cause M*, by $P_L$ it must be part of a process consisting of direct cause-effect pairs, each of which does some of the causal work required for the next. But causal work is done only once, and all the causal work to be done here is done by the events in the process from P to P*. Put simply, there is *no way* for M to be process-linked to M* unless it *just is* one of the links in the process from P to P*. We can think of it in the following metaphoric terms: M is trying to find some work to do, but there are no gaps in the process from P to P* for M to fill. Whichever way M looks, all the work is already done by something else. And so if M causes M*, then (given that M and P are simultaneous) it follows that M=P. Again, notice that in the above derivation, $P_L$ figures as a *necessary* condition on causation. We may now write down the premises behind the causal exclusion problem. Following Crane [1995], I will write them as a mutually inconsistent set of propositions.[158] Putting the matter this way has the advantage of making transparent the taxonomy of possible solutions to the problem.

| | |
|---|---|
| $E_M$ | Mental events cause behaviours. |
| $C_P$ | Every physical event *y* that has a sufficient cause at *t*, has a complete, sufficient *physical* cause *x* at *t*. |
| ¬ID | Mental events are not identical to physical events. |
| $P_L$ | An event *x* is a (sufficient) cause of an event *y* iff *x* and *y* are parts of a (sufficient) causal process in which *x* occurs prior to *y*. |
| CW | The causal work required for a given effect is done at most once. |
| $T_{CW}$ | If *x* is a direct sufficient cause of *y* and *y* is synchronically sufficient for *z*, then *x* does all the causal work required for the occurrence of *z*. |

[158] Crane formulates the problem for Davidson events causally related in virtue of their properties; my treatment, of course, is in terms of causally related property-instances. There are other significant differences to Crane's version, and I do not attribute the version I give to him. In particular, Crane runs the exclusion argument in terms of a principle of non-overdetermination, which he understands (as I do) as a ban on massive coincidence. Crane accepts that the efficacy of supervenient causes would not represent such coincidence, but thinks it must involve denial of a principle he terms the *homogeneity of mental and physical causation*. I will return to the issue of homogeneity in 5.4. See Crane [1995] p.229 for details.

I omit $G_C$ and $G_C$' from this formulation for simplicity – nothing turns on this omission as $P_L$ presupposes $G_C$ and $G_C$' This is because the generativity principles, along with the definition of a causal process, serve to define the terms of $P_L$. The six principles above are jointly inconsistent. Replacing $P_L$, CW and $T_{CW}$ with $O_D{}^4$ yields an argument of similar form to the causal argument, with $O_D{}^4$ replacing the more liberal coincidence-based principle of non-overdetermination we employed. So formulated, we would have four inconsistent and independent theses, any three of which can be taken as premises in an argument for the negation of the remaining one. Stating the connection between causal exclusion as a problem on the one hand, and an argument on the other, in this way, is not new. It is implicit in Crane [1995] and explicit in Sturgeon [1998]. Sturgeon lists four inconsistent theses, which are, in essence, my $E_M$, $C_P$ and ¬ID above, along with a principle of non-overdetermination.[159] And correspondingly, Sturgeon is able to generate four exclusion arguments, which involve endorsing three of the inconsistent theses as premises, yielding the negation of the other as a conclusion.[160] For the time being, then, I will consider that we have *six* arguments here, each formed by endorsing the other five propositions as premises in order to deny either (i) $E_M$, (ii) $C_P$, (iii) ¬ID, (iv) $P_L$ (v) CW, (vi) $T_{CW}$.

The causal exclusion problem, when formulated (for instance) in Sturgeon's terms, has seemed to many to be intractable, because each of the possible arguments yields the denial of a well-supported or else intuitively highly plausible thesis. For my part, I think that a great deal of this intractability stems from the fact that the problem is consistently stated in terms of a ban on overdetermination, the denial of which has been standardly considered too implausible to countenance. However, it is the denial of $O_D{}^3$ that is absurd; denying $O_D{}^4$ patently is not, as $\neg O_D{}^4$ does not entail widespread

[159] I do not know whether Sturgeon would endorse $O_D{}^4$; but it seems clear he *would* endorse a coincidence based principle such as $O_D{}^3$. See Sturgeon [1998] pp.413-4.
[160] Horgan [2001] endorses alternative versions of Sturgeon's four theses, along with the premise that mental properties are 'real', and correspondingly is able to generate not four but *five* arguments, the extra argument being one in favour of eliminativism. I am not convinced that mental property realism is not already implicit in $E_M$, for on the view of events I endorse, if there are no mental properties, it is unclear that there are any mental *events* either.

coincidence. Casting the problem in terms of the principles that *underpin* $O_D$[4] makes our task appear far more tractable. In what follows, we will consider which of the six arguments supported by our present formulation of the problem is most plausible. This task is easier than it seems, for it is a relatively straightforward matter to dismiss those that entail ¬CW and $\neg T_{CW}$, leaving us the task of arbitrating between arguments for (i) $\neg E_M$, (ii) $\neg C_P$, (iii) ID, and (iv) $\neg P_L$.[161] My overall strategy for the remainder of this chapter is to show that argument (iv) is by some considerable distance the most plausible of the four. In the next section, I briefly consider some extant denials of $E_M$, $C_P$, ¬ID and $P_L$. I will suggest that any one of (i)-(iv) can be made plausible by the severity of the exclusion problem – if, for instance, $\neg E_M$ is the only way to solve the problem, then $\neg E_M$! My treatment will be brief because in 5.5, I argue that there are clear counterexamples to $P_L$, and that we should be in no doubt whatever about the best way to solve the exclusion problem – all we need to do is reject the process-linkage account, which is highly dubious on independent grounds. As such, the remainder of this chapter ought to be seen merely as setting out the logical geography of responses to the exclusion problem.

## 5.4. A brief taxonomy of solutions

1. Denying CW

¬CW makes the problem of finding causal work for M to do goes away – it can simply do the same work as P does, over again. But denying CW is implausible on general metaphysical grounds. Causal work, on the present conception, involves the transfer of conserved quantities. This being the case, it is arguably not *possible* to do the causal work necessary for a given event to be done twice. This is because the conserved quantities in question are quantifiable – "doing the work twice" is really just doing twice the work. Suppose two builders need to move a large stone, and move it together to its new location. The causal work here is not done twice; rather, it

[161] Arguments (i), (ii) and (iii) have been endorsed in the literature, in various forms, as solutions to the exclusion problem. Explicit discussion of the nature of 'causal work' is quite thin on the ground; as such, so are explicit denials of process view of causation as solutions to the causal exclusion problem. There are, however, authors who implicitly deny the process account, by showing that on *their* account of causation, the problem does not arise. More on such theories is to follow.

is done *once* by *two* builders, each of whom does half the work. On the assumption that causal work involves the transfer of conserved quantities, it is simply not possible to do the same work twice.

2. Denying $T_{CW}$

We follow up the dismissal of (v) with an equally summary dismissal of argument (vi). $\neg T_{CW}$ makes the problem of finding causal work for M to do go away – for now there is *extra* work to be done in causing M* that is not done by any of the causes of P*. However, if $T_{CW}$ is false, then when you build a wall by laying the bricks in appropriate places, your work is not yet done – for in addition to producing *this* particular aggregate of bricks, you have to do the work of making it so that there is a *wall* where the aggregate is! In short, denying $T_{CW}$ is inconsistent with the view that in doing the causal work necessary for the occurrence of an effect *y*, we thereby do the causal work necessary for all effects that are nothing over and above *y*. I note in passing that there may be relations of synchronic sufficiency according to which $T_{CW}$ comes out false – namely the sort of simultaneous causal relations that Lowe endorses.[162] In those cases, clearly all the causal work isn't done merely by causing the synchronically sufficient event. However, $E_M$ tells us that mental events cause *behaviours*, and these latter we know are not synchronically *caused* by physical events. As such, even if $T_{CW}$ is not true for all synchronic sufficiency relations, it is clearly true in the cases that matter to us.

3. Denying $E_M$

Denying $E_M$ solves the problem of causal exclusion by biting the bullet, and accepting that mental events are inefficacious on the grounds that they do no causal work. I will mention two theories that can plausibly, although not uncontroversially, be grouped together as denying $E_M$. The controversy stems from the fact that they can also be seen as denying what Crane terms the 'homogeneity of mental and physical causation'.[163] I will return to this point presently; what we can all agree upon is that

[162] Lowe [2000]. See 1.4, 3.2 and 3.3 of this work for brief discussion.
[163] See Crane [1995] pp.229-33.

the theories to be described deny $E_M$ *on the assumption that the process account of causation is true*. Into this category, we can place Kim [1984]. Kim once believed in a thing called 'Supervenient Causation' which obtains between supervenient properties whose supervenience bases cause each other.[164] Supervenient causation enables us to hold on to $C_P$, ¬ID, and the process account via the thought that only subvenient causation is causation proper. Physical properties do all the causal work; supervenient mental properties 'cause' solely in virtue of the causal relationship between their subvening properties, for as we have seen there is no causal work left over for them to do. Supervenient "causation", at least for Kim [1998], is, as we saw in 5.2, a pseudo-process "like the series of shadows cast by a moving car…." Given the process view of causation, then, supervenient causation is perhaps best seen as a form of epiphenomenalism.

Jackson and Pettit [1990] offer a broadly similar account. They accept that the exclusion problem entails $\neg E_M$, and present a phenomena-saving account of how it is possible to give causal explanations in terms of inefficacious properties. Jackson and Pettit distinguish process from *program* explanations. Program explanations work because instances of the properties they cite ensure (or at least make it significantly probable) that a process of a certain kind occurs. For instance, the fragility of a vase causally explains its breaking despite the fact that dispositional properties do not do any causal work. This fact is then explained via the thought that fragility programs for its categorical base properties, instances of which *do* figure in the causal process that causes the breakage. Mental properties, according to Jackson and Pettit, are causally *relevant* without being causally *efficacious*. The central burden of such a theory is that if the process account is true for physical causation, and mental causal relevance supervenes on it, then an explanation is needed as to what distinguishes the *genuine* relevance of mental properties from the correlations between supervenient but causally *irrelevant* properties, like 'shadows cast by a moving car'. The relevance of mental events cannot, of course, be distinguished from the irrelevance of shadows by

[164] See Kim [1984] for details of this theory, which Kim no longer endorses.

dint of an appeal to causal work – rather, we have to rely on things like explanation, counterfactual dependency, and laws.

Giving such an account has not proved an easy task. On Jackson and Pettit's view, for instance, it is not sufficient for causal relevance that a property ensures that a process occurs. The programming property must also figure in an explanation that carries modal information not carried by the process explanation. In order to do this, programming causes must exhibit "invariance of effect under variation of realization," ([1990b] p.202). Call this 'realizer-invariance' for short. All instances of a property such as temperature instantiate the same thermodynamic laws whatever their microphysical realizations, and so temperature is a genuinely relevant property. But now suppose that all and only things at temperature T emit a characteristic red glow. It looks as though explanations citing the glow will be just as realizer-invariant as those citing T, for the glow by hypothesis will not depend on how temperature T is realized in any particular case; and the glow will clearly program for the efficacious microphysical properties that figure in the corresponding process explanation. But problematically, the glow seems a clear cut example of a causally *irrelevant* property. I should note in passing that I agree with Jackson and Pettit that realizer-invariance is an important feature of causal explanations involving realized (or more generally, *supervenient*) properties – indeed, in 6.2, I appeal to this very fact to explain why supervenient properties are explanatorily non-redundant. My point here is simply that it is less than clear that realizer-invariance provides a *sufficient condition* for causal relevance.

I reiterate at this point what I said in 3.1, that given the weight of *prima facie* evidence in favour of $E_M$, anyone wishing to deny it had better have run out of plausible alternatives. Since the evidence is only *prima facie*, however, it is defeasible by (*inter alia*) the absence of any alternative solutions to the exclusion problem. If, in the end, there is no way to solve the problem other than giving up on mental causation, then so be it. Now as I mentioned, the accounts described above can also be taken to be denying (not $E_M$, but) a further assumption, namely that mental and

physical causation are 'homogeneous'. As will I point out presently, this assumption is implicit in $P_L$, so I group those who deny homogeneity together as endorsing $\neg P_L$, of which more below.

4. Denying $C_P$

Denying $C_P$ solves the problem by allowing us to consider the mental and physical causes of a behaviour as *jointly* causally sufficient for it. This is probably the least popular of all the responses to the causal exclusion problem, probably because denying $C_P$ is typically associated with positions such as emergentism and dualism. These positions are unpopular for two reasons. First, although, for instance, emergentism is coherent, there is arguably no reason to think that it is *true*.[165] If particles were accelerated without the need for particle accelerators, or muscles contracted without any preceding brain activity, then we would have reason to actively doubt $C_P$. The phenomena seemingly aren't like that though, and the second reason $\neg C_P$ is unpopular is that as we saw in 3.2, it is possible to argue *for* $C_P$ on that very basis. Set aside for the moment the question whether the argument is any good; what I take to be relatively uncontroversial is that there is no evidence *against* $C_P$.

Cartwright is often interpreted as denying $C_P$ (and she may well deny it) but in fact her arguments only support the weaker claim that we have no reason to believe it is true. It is worth taking a moment to see why this is so. Cartwright argues that there is no reason to suppose the motion of a leaf on the wind is governed by physical laws.[166] The reason, roughly, is that physical laws are tested against very specific and carefully controlled background conditions. Experimenters testing a force law for two charged particles will take great care to screen off any external forces that may affect the way the particles move. The result is that the law in question is confirmed (or disconfirmed) only relative to the model in which it was tested, and so we are epistemically justified in generalising it only to relevantly similar situations. Surely,

[165] In chapter 7 I will agree with this, but claim that given the currently available evidence, the mere *possibility* of emergence, as detailed in chapter 6, is enough to render $C_P$ empirically unsupported.
[166] Cartwright [1994] pp.234-5

one might object, the air molecules whose motion composes the wind act on the leaf by exerting forces on it? Cartwright denies there is any reason to believe this either, as forces, she claims, also belong to the specific models within which force laws are tested.

I will make only two very brief points in response. First, Cartwright's argument is an epistemic one against the justification for $C_P$, and as such provides no reason to think that $C_P$ is *false*. The fact that we test laws against specific models does not entail that the laws are not universal. Loewer makes a similar point, claiming that all Cartwright has done to undermine $C_P$ is point to the *possibility* that regularities that hold under laboratory conditions might fail under less carefully controlled conditions.[167] It is, as Loewer points out, quite another thing to have a *reason* for thinking that regularities that hold in the laboratory *do not* hold outside of it. Second, I would tentatively ask: surely one of the central marks of a good scientific theory is that it generates novel predictions that turn out to be true; but what is a novel prediction if not one that goes beyond the model against which the predicting laws have been tested? If that is true, then we have reason to believe that there are some regularities that hold both inside *and* outside of the laboratory. Denying $C_P$ is consistent, but there is no independent reason to do so. Still, if denying $C_P$ proved to be the only way of solving the exclusion problem, then that in itself would count as just the sort of reason we lack.

5. Denying ¬ID

This solution is the one favoured by Kim, and (obviously) solves the exclusion problem by 'allowing' mental events to do exactly the same causal work as physical events. As we saw in 2.3, on the assumption that mental properties are multiply realized, identifying mental property-instances with physical property-instances results in eliminativism. Now I am considerably in sympathy with those tempted to object at this point that the conjunction of multiple realization and ¬eliminativism is at least as plausible as any of the other premises of the exclusion argument. If that is

[167] In his [2001a] pp.52-3.

true, then ¬ID really ought to be the last premise we reject, for if we have independent reason for endorsing any of the premises, it is surely this one. In addition, it seems that the exclusion argument for ID will *generalize* to show that *all* efficacious properties are identical to properties of basic physics; after all, nothing in the present formulation depends on any *particular* relation of non-identity between mental and physical events.[168] The problem is that anything that *isn't* identical to the physical will be deprived of causal work by the sufficient physical causes, unless it is *identical* to those causes. But now anything that is multiply realized by the physical will be eliminated, and that makes the elimination everyone's problem.

Worse than that, though, what if it turns out that it's multiple realization *all the way down* without end? As Schaffer [2003] points out, that there is no fundamental level is an open empirical possibility, given current evidence. But if the efficacy of *every* property is excluded by the efficacy of another, then there is no efficacy at all – this is the problem of causal drainage.[169] And if *every* property-instance must be identified with an instance of one of its realizers, then it's elimination all the way down too – call this the problem of *property*-drainage. The drainage problems suggest an obvious *reductio*: "it's an open empirical possibility whether or not there is a fundamental level; it's not an open empirical possibility that there are neither any properties nor any causation; therefore the existence of properties and causation does not depend on the existence of a fundamental level. But if the exclusion argument for ID is sound, the existence of properties and causation depends on the existence of a fundamental level. Therefore the argument is not sound." Of course, Kim must agree with the objector that the very existence of causation and properties can't depend on there being a fundamental level; what he denies is the conditional, *if* the exclusion argument

[168] I am calling this relation 'realization' here, but I intend this in the broad sense described in our discussion of the multiple realizability of temperature in 2.3, according to which the specifications that define the second-order realized properties need not be causal.

[169] See Block [2003] for an account of the problem, and an attempt at reconstructing Kim's solution, to be found in Kim [1998] pp.84-7, pp.116-8, and Kim [2003]. I understood little of Kim's solution until I read Block; my summary treatment here is attributable to his [2003], except where I support Kim below in denying the causal drainage problem. I follow Block in taking Kim's remarks of [1998] pp.84-7 concerning the distinction between levels and orders to be irrelevant to the problem at hand, and describe instead only the much more plausible response Kim (arguably) gives at pp.116-8, and which Kim explicitly endorses in his [2003] reply to Block.

in sound, then the existence of causation and properties depends on there being such a level. It is worth taking a moment to see how Kim's denial of this conditional works.

Kim has two related reasons for denying that the exclusion argument makes the existence of causation and properties dependent on the existence of a fundamental level. Both strategies have a certain plausibility, and can be found in Kim [2003]. The *first strategy*, in very rough outline, consists in the claim that although there may not be a fundamental level in the sense of there existing any non-composite particulars, there is no multiple realization between 'levels' understood as ordered by microphysical mereology. An instance of the property of being $H_2O$ is identical to a specific mereological configuration of atoms; the atoms in turn are identical to specific mereological configurations of subatomic particles, and so on indefinitely if there are no mereological atoms. The situation is importantly different to that of thermodynamics, where no particular structural property is necessary in an aggregate of molecules for it to be at a given temperature.[170] However, Kim claims, the property of being $H_2O$ is identical to a structural property of atoms, which in turn are identical to structural properties of quarks, and so on all the way down. If there is endless mereological decomposition, then that merely reflects the fact that certain microphysical properties are *infinitely structural*, and not that every such property is realized by a distinct and even more microphysical property. Thus the efficacy of instances of microphysical properties is not excluded by the efficacy of the property-instances at the next mereological level down, for these 'two' instances are in fact one.

A response to the first strategy will lead us nicely into the second. Response: isn't it also an open empirical question whether microphysical properties are multiply realized at the next mereological level down? Might not the property of being a quark, say, be like temperature in this respect? I suppose it might be argued that even a given

[170] I do not intend by this comparison to attribute my views on thermodynamics to Kim. As we saw in 2.3, Kim would maintain that thermodynamic properties are eliminated by functional reduction and the causal inheritance principle.

chemical element will, on different occasions, be realized by different quantum states of the subatomic particles that compose it. The *second strategy* is, to my mind, a knock-down argument of the drainage problem, and it relies on the fact that the exclusion argument depends on the completeness of physics. If the quantum level is multiply realized at the next mereological level down Q-, then arguably quantum physics will be causally *incomplete* due to the possibility of quantum events having Q- causes but no quantum causes. If Q- is causally complete, then causal efficacy and properties will drain down to the Q- level; but the crucial thing is that the drainage will not go down any further, for there will be no level with respect to which Q- is causally *in*complete to exclude the efficacy of Q- causes. This is closely related to the first strategy, in the following way: if Q- is causally complete, then it will not be multiply realized. But then Q- properties will be identical to structural properties of Q--, and so on all the way down. If, on the other hand, it really is *multiple* realization all the way down, then no level will be causally complete, and so the causal exclusion problem cannot get off the ground. Either way, there is no problem of property or causal drainage. Neither strategy, of course, prevents the efficacy of all *other* properties (and so too their reality, given the eliminative consequences of combining instance-identity with multiple realization) draining down to the first properties that are not multiply realized. Physiological and thermodynamic properties, for instance, will not survive; certain basic chemical properties (such as the property of being $H_2O$) plausibly will. I take it that Kim is more likely to see this parsimonious consequence as a *virtue* of his theory than a vice.

One can, of course, be an identity theorist without being an eliminativist, by endorsing alternative metaphysics of the things being identified. One such position, we have already encountered in 2.3. If Shoemaker is right that properties are sets of conditional powers, then we can identify mental property-instances with *parts* of physical property instances. Correspondingly, this partial endorsement of ID solves the exclusion problem via the claim that mental events do *part* of the causal work

required for their behavioural effects.[171] Notice that the strong Shoemakerian metaphysic, according to which properties are sets of causal powers, is required for this identification to go through – on the weaker (and correspondingly, far less controversial) view that properties *confer* sets of causal powers on their bearers but are not *identical* to such sets, the efficacy of mental properties will once again be excluded. For if the powers conferred by mental properties are subsets of those conferred by physical properties, then the causal powers of a mental property-instance will be identical to a subset of the powers of a distinct physical property-instance. Which, if anything, makes it *even more* transparent that mental property-instances are shorn of causal work by their physical realizers. The trouble now is that identifying properties with sets of causal powers brings with it some heavy-duty metaphysical baggage. For instance, it entails that the laws of nature are necessarily true – for properties could not behave in a different way and yet remain the same properties. And relatedly, it entails that all causes metaphysically necessitate their effects – a position many would consider untenable, given the views (i) that cause and effect are distinct existences, and (ii) that there are no metaphysically necessary connections between distinct existences.[172]

A second brand of identity that avoids eliminativism is *trope theory*.[173] According to this view, the relata of causation are not property-instances conceived as structured particulars comprising individuals, properties and times, but *particularised properties*, thought of as, for instance, as something like '*this* yellowness'. Properties, on this account, are classes of tropes related by relations of resemblance or similarity. Now this metaphysic invites the view, endorsed in Robb [1997], that $E_M$ and $C_P$ are claims

[171] Or alternatively, and perhaps better, they do those parts of the causal work that *are* required for their behavioural effects. The extra bits of work that physical events do, although required for particular realizations of behaviour, and not required for the behaviours themselves. See 2.3 for discussion of the related subset account of realization.

[172] The problem mentioned here (to the extent that it is a problem at all) is a problem for functionalism about any domain of properties. Any property whose individuation involves having an effect will be one whose instances are metaphysically connected to instances of the effect. On the Shoemaker account, this is a problem not just for functionalists, but for all properties, since in effect, Shoemaker holds that *all* properties are functional (though not necessarily *second-order*). These matters are beyond the scope of the present work.

[173] See Robb [1997]; Ehring [1999] for discussion of the application of trope theory to the problem of causal exclusion.

about mental *tropes*, while ¬ID is a claim about mental and physical *properties*. And identifying tropes does not force us to identify the similarity classes of which they are members, for a single individual can be a member of many such classes. Trope theory promises an elegant solution to the exclusion problem.[174] However, it too has quite a lot of baggage, most of which concerns its appeal to relations between tropes to give an account of properties. How is resemblance to be understood, if not in terms of resemblance in respect of a shared *property*? Trope theorists must avail themselves of primitive resemblance relations to unify the tropes into classes. But how then are resemblance relations to be distinguished from other (i.e. causal) relations? We seem to need *another* resemblance relation that takes relations as relata. It is not clear to me whether this regress is either infinite or vicious, but either way it makes trope theory look anything but elegant as a general metaphysic of properties. Still, if ID is the only way to solve the exclusion problem, then trope theory, like Shoemaker's causal powers account, might recommend itself as among the most plausible ways to preserve realism about nonphysical properties.

6. Denying $P_L$

This strategy is implicit in several distinct responses to the exclusion problem. These approaches can be thought of as falling into two broad categories: those that deny 'homogeneity', and those that endorse an alternative theory of causation. The two approaches are intimately related – the first involves accepting that something like a process-linkage account is correct for physical causation, but denying that it provides the correct model for *mental* causation. The second involves endorsing a general theory of causation and showing that given the theory, there is no problem of causal exclusion; such a theory must be inconsistent with $P_L$ and so (by implication) this approach involves denying the process-linkage account. For Crane, the 'homogeneity of mental and physical causation' means homogeneity of the *concept* of causation

[174] Noordhof [1998] argues that tropes in fact only relocate the problem, for there are clear cases where *x*'s being a trope of P seems less causally relevant than *x*'s being a trope of Q. For instance, Yablo's pigeon, conditioned to peck at red things, fails to peck at coloured things unless they are coloured red (Yablo [1992]). The redness of an apple seems clearly more relevant to a particular peck than does its colouredness. Trope theory, however, struggles to accommodate such distinctions, for it seems that nothing has the effects it does in virtue of being a member of a similarity class.

employed in $E_M$ and $C_P$.[175] Notice that this assumption is implicit in our principles of causal work – $P_L$ is taken as a necessary and sufficient condition for causation *simpliciter*, not any particular variety thereof. It might be suggested, however, that since we have conceived causal work in terms of *physical* work, the theory of 5.2 is most suitable as a theory of *physical causation*. Following this line of thought, we ought then to reformulate $P_L$ in terms of physical causation. Given the further assumption of homogeneity, this new principle will then apply to the causes premised by $E_M$, and the argument goes through exactly as before.

Crane takes homogeneity to be an essential part of *any* causal argument for physicalism, for such arguments have it in common, he claims, that they all motivate physicalism via "*conflict* between mental causation and the completeness of physics [my italics]."[176] I am unable to agree, as I take the argument I presented in 3.4 to motivate physicalism by appeal to *coincidences*. As I argued in 3.3, and again in 4.4, generating the required coincidences does not require that mental and physical causes are related in the same way to any particular effect. For instance, in 4.4, we saw that the fact that the behavioural effects of a mental cause always have sufficient physical inducers stands in need of explanation just as much as if the mental and physical causes were related to behavioural effects in exactly the same way. Homogeneity is certainly required in the *exclusion argument*, however, and the reason is simple. If mental causes do not cause their effects by doing causal work, then quite plainly it does not count against $E_M$ if the conjunction of $C_P$, ¬ID and $P_L$ leaves no causal work left for mental events to do.

Of course, denying homogeneity leaves room for supervenient properties and programmers to be genuine causes after all. Once it is granted that mental causation is of a different *kind* to physical causation, we are free to maintain that program explanations *do* cite genuine causes, but that these causes aren't the same *kind* of causes cited in process explanations; or that supervenient causation is genuine

[175] See his [1995] p.219 and pp.232-5.
[176] Crane [1995] p.235.

causation, of a different kind to *sub*venient causation. Such accounts will still face the problem outlined in (3) above, viz. that it is not easy to articulate a counterfactual or law-based criterion that adequately distinguishes genuine causes from epiphenomena. An appeal to causal work would solve this problem, but of course no such appeal is available. It is difficult to see how to arbitrate between the view that denying homogeneity describe mental properties as epiphenomena, and the view that they describe them as causes of a different kind. Notice, however, that inhomogeneity accounts provide an obvious rejoinder to proponents of $P_L$ – for they might simply insist that process-linkage is causation proper, and that merely *calling* a relation other than process-linkage 'type-X causation' doesn't *make it a genuinely causal relation*. This, I take it, is one of the central reasons why Kim no longer endorses supervenient causation. I am in agreement with Crane that an "exchange of intuitions about what exactly 'epiphenomenon' means" would be fruitless.[177] We should keep the possibility of the inhomogeneity of mental and physical causation in mind – indeed, as I mention in 5.5, there are those who take certain difficulties in the analysis of causation to suggest that the concept of causation itself (regardless of domain) is multifarious. If this is true, then indeed mental and physical causation may be inhomogeneous, not because different concepts of causation apply in the mental and physical case, but because the concept that does apply itself picks out different relations depending on context.

The process account can be endorsed as a theory of physical causation and denied as a theory of mental causation; it can also be denied as a theory of any kind of causation, which is how I read Yablo [1992].[178] Crane, however, attributes a denial of homogeneity to Yablo.[179] Given Yablo's reliance on counterfactuals, this is understandable, for appealing to counterfactuals is a common way to try to account for the causal relevance or efficacy of states that are not causes in the same way as physical causes. Now I do not know whether Yablo would agree that his theory makes

[177] Crane [1995] p.234.
[178] See 4.3 for an account of Yablo's theory of causation.
[179] In his [1995] p.234.

physical and mental causation inhomogeneous; but the reason I do not know this is that I do not know whether or not Yablo holds that the process account (or anything like it) is true for *physical* causation. What I do know is that if the process-linkage account is rejected, then pace Crane, accounts such as Yablo's do not have to deny homogeneity. Refer back to the causal diagram at the beginning of 5.3. It is perfectly consistent to maintain that M causes M*, P causes P*, but that neither M nor P does the causal work necessary for M* or P* respectively. How so? M's causing M* consists, for Yablo, in counterfactual dependence plus proportionality; but nothing prevents us from holding the same about P's causing P*. It is only if we insist *in addition* that P causes P* by virtue of a causal process that we have to accept that M's relation to M* can't be the same as P's relation to P*. But why insist on that? It seems quite natural for Yablo to maintain that causation consists in counterfactual dependency plus proportionality, *and nothing else*. It is of course open to Yablo to maintain that in addition to the other requirements, physical causes must also be process-linked to their effects. My point, though, is that an additional subscription to the process account for physical causation is independent of the counterfactual and proportionality elements of Yablo's theory.

Once the process account is denied even for physical causation, nothing stands in the way of the relation that holds between M and M* being just the same relation as the relation that holds between P and P*. On this view it is the *relata* that differ between mental and physical causation; the *relation itself* is the same. Notice that this makes the task of distinguishing real causes from epiphenomena without appealing to causal work everyone's problem, and not just a problem for those wishing to give an account of *mental* causation. It is of course an open question whether any such account can be given; however if Yablo's account is successful (and I will not here attempt to address the question whether or not it is) then the exclusion problem is solved without the need to deny homogeneity. Counterfactual dependency plus proportionality between physical causes and their effects, clearly does not exclude the very same relations holding between the mental causes and *their* effects.

Horgan has recently advocated a theory of causation not dissimilar to Yablo's.[180] According to Horgan, the concepts of causation and causal explanation have implicit 'level-parameters'. He seems to argue as follows. First, he claims that causal explanation involves subsumption of events under appropriate counterfactual-supporting generalizations. Second, he argues that different ontological levels are characterised by different such generalizations – the appropriate generalizations for explaining a behaviour will be psychological in character, whereas those appropriate for explaining a physical event will be physical. Third, he appears to endorse the view, shared by Baker [1993], that the concept of causation is not separable from the concept of causal *explanation*. But from these theses it follows that there will be many levels of causation, which complement, but do not compete with, each other. There is no exclusion problem because mental and physical causal explanations have different *explananda*. Since the level-parameters are not explicit in claims of causal efficacy, we are apt to treat efficacy as efficacy *simpliciter*. This, Horgan claims, is a mistake – the exclusion problem is a 'cognitive illusion' that occurs when we fail to fully understand what it is that causal claims claim. The precise details are unimportant for my present purposes. What matters is that in indexing the truth of causal claims to causal explanatory claims, and these latter to counterfactuals, Horgan denies $P_L$, for as I take it is by now familiar, process-linkage between events *x* and *y* is by no means a necessary condition for *x* and *y* to instantiate a counterfactual-supporting generalization. In addition, I can discern nothing in Horgan's views that might count as a denial of homogeneity – for he endorses just the same criteria for the causal-explanatory relevance of both mental and physical properties, relative to their proper levels.

Finally, Menzies [2003] endorses a theory of causation that is similar to those of Horgan and Yablo in maintaining that causation is an intra-level relation. Menzies goes further, however, in explicitly arguing that his account is consistent with

[180] See his [2001], which develops themes introduced in his [1997].

homogeneity. The account he gives is as follows. First, he defines a process as a temporally ordered sequence of events, and holds that:

> The counterfactual dependence of E on C relative to the model X *picks out a process*…if and only if the process is present in all the most similar C-worlds generated by the model that are E-worlds and is absent in all the most similar ¬C-worlds generated by the model that are ¬E-worlds.

Causation is then defined by Menzies as follows:

> C is a *cause* of the distinct state E relative to the model X of an actual situation if and only if
>
> 1. E counterfactually depends on C relative to the model;
> 2. this counterfactual dependency picks out a process;
> 3. this process connects C and E in the actual situation.[181]

The idea here is that relative to a physiological explanatory model, a behaviour will depend counterfactually on physiological causes; similarly *mutatis mutandis* for a psychological model. Crucially, for Menzies, the counterfactual dependency of behaviour on mental causes will pick out a psychological process, and its dependency on physiological causes will pick out a *different*, *physiological* process.[182] Call the behaviour B, and let it depend counterfactually on mental cause M. On the psychological model, given multiple realization, the set of closest M-worlds that are B-worlds will include worlds at which the physiological process that actually realizes M does not occur. There is no exclusion problem, on this view, because $C_P$ and $E_M$ are indexed to different models, and so the relevant counterfactual dependencies will pick out different processes. Although Menzies appeals to the notion of a process, he detaches this from any account of how the events in the process are *linked*. In particular, there is no mention of causal work – distinct processes can coexist within different models, as there is no supposition that the events in the process do the causal work necessary for the events that follow them. Thus Menzies too (implicitly) denies

[181] See Menzies [2003] p.212 for the definitions quoted here. Care must be taken not to confuse Menzies' talk of process with the process-linkage account. No flow of conserved quantities is necessary, for Menzies, for a sequence of events to count as a process.
[182] See his [2003] pp.215-23 for the detailed application of his theory to the exclusion problem.

$P_L$. He endorses homogeneity, as the concept of causation defined above is applied equally to both mental and physical causes. That is, his theory contains no admission that physical processes are somehow different, or that the events that form a physical process are linked in a stronger way than those of a process defined by a non-physical model.

The accounts I have grouped together as denying $P_L$ all have a common feature, which is that they all seek to define causation in terms of extrinsic relations between events. On Yablo's account, for instance, causation between actual events involves facts about what goes on at other worlds; on Horgan's account, it involves the events instantiating counterfactual-supporting generalizations. These accounts, as I will explain in the next section, can be further grouped together as *probability* accounts of causation; we turn now to the question whether causation is most plausibly analysed in terms of probability or process. It is widely accepted that no extant theory of causation accommodates all the relevant causal phenomena. However, I will argue in what follows that process accounts are doomed to failure in a way that probability accounts are not. For while probability accounts at least offer the promise of being able to accommodate problem cases, process accounts are forced to deny that certain problem cases are cases of causation at all.

### 5.5. Causation: probability or process?

My argument in this section will be brief. Much of what I have to say more or less reiterates Schaffer [2000], and especially [2001]. Following Schaffer I distinguish two broad categories of theories of causation, viz. *probability* and *process* accounts. The process account we have already examined (at least in one form) in 5.2-5.4 above, and it is $P_L$ to which I will refer when, in what follows, I speak of process accounts. Among probability accounts are the various regularity and counterfactual theories according to which causation is a matter of the right dependency relationships between events, regardless of how, if at all, the events are physically *connected*. The reason these accounts can be grouped under the heading of probability accounts is that they are more-or-less convergent for indeterministic causation. In this case, law-based

accounts, for instance, will hold that causation involves a lawful regularity between the occurrence of the cause and an increased probability of occurrence of the effect; counterfactual theories will hold that if the cause had not occurred, then the probability of the effect's occurrence would have been less. I will first briefly discuss problems for each style of account, and point to the manner in which each can be seen as feeding off the weaknesses of the other. I suggest, also following Schaffer, that if there is to be an account of causation that accommodates *all* the relevant phenomena, then it *must* include an appeal to probability – no "pure" process-linkage account such as $P_L$ will suffice, for there are just too many clear-cut cases of causation that such accounts are *in principle* unable to cover. My own contribution will be to suggest that once this much is admitted, the exclusion problem disappears, for it is only on a "pure" process account that the problem arises. The exclusion problem depends on process-linkage as a *necessary condition* for causation; but this is precisely the claim that renders the account subject to counterexamples, and requires the incorporation of elements of the probability view.

Now in 5.4, we saw that there are several probability accounts according to which the exclusion problem does not arise. For instance, Yablo's proportionality account, Menzies' counterfactual account, and Horgan's causal explanatory account. In general, there just isn't anything peculiar in the thought that, for instance, both M and P raise the probability of M*, perhaps to differing degrees; or that M* depends counterfactually on both M and P. Construing causation in probabilistic terms invites what Horgan terms 'causal compatibilism' – the view that mental and physical causation do not exclude each other, and maybe (as Yablo thinks) even *complement* each other.[183] Conserved quantities such as energy and momentum, however, are importantly different. If two causes transfer the same amount of energy to an effect, then it receives double the amount transferred by each cause. Two billiard balls can not transfer *the very same energy* to a single ball with which they both collide.

[183] Horgan [1998], [2001]; Yablo [1992]. Similar themes are to be found in Jackson and Pettit [1992] and Sober [1999], who argue that mental and physical explanations complement rather then exclude each other. We return to this matter in our discussion of novelty and redundancy in 6.2.

Process accounts of causation, it seems, entail that "real" causation is only done once. For the purposes of the argument to follow, I will consider only specific elements of probability and process accounts. In particular, the claims with which I want to take issue are (i) that raising the probability of an effect is a necessary condition for causing it; and (ii) that process-linkage is a necessary condition for causation. As we shall see, there are clear counterexamples to each of these claims. I will mention in passing similar counterexamples to the corresponding sufficiency claims.

The relationship between probability and process accounts is interesting – there is a sense in which each can be seen to thrive off the failings of the other. It is widely accepted, for instance, that "late pre-emption" is a serious problem for probability accounts.[184] Late pre-emption is a species of redundant causation in which the actual (pre-empting) cause of an effect cuts off another event that would have caused it in the absence of the actual cause. In cases of late pre-emption, the pre-empted cause is typically prevented from causing the effect by the fact that the process linking the *actual* cause to its effect has gone to completion. By way of illustration, consider the following case, described in Lewis [2000]. Billy and Suzy both throw rocks at a bottle. Suzy's rock arrives fractionally before Billy's rock, shattering the bottle, and Billy's rock, we may suppose, makes no difference at all to the shattering, passing through the empty space where the bottle used to be. Suzy's throw is the pre-empting cause, Billy's the pre-empted cause. The trouble this scenario raises for probability accounts is that Suzy's throw does not now raise the probability of the bottle's shattering. Billy's pre-empted throw, we may suppose, *guarantees* that the shattering will occur anyway, even if Suzy's throw misses. In addition, suppose for the sake of argument that Billy's rock is twice the size of Suzy's. Suppose further that the impact of Suzy's rock is *only just* enough to break the bottle in the actual world; in very close possible worlds the impact of her rock suffices only to move the bottle out of the way of Billy rock. Billy's rock, on the other hand – had it impacted – would have been more than enough to cause the shattering. In the situation described, Suzy's throw

[184] See Lewis [1986d] for detailed discussion of various types of preemption.

actually *lowers* the probability of the shattering. Uncontroversially, however, Suzy's throw *causes* the shattering, and so probability raising cannot be a necessary condition for causation. In passing, we may note that the situation described also provides a counterexample to the claim that probability-raising is *sufficient* for causation. Suppose Suzy is a bad throw, so that there is a significant probability that her rock will miss the bottle altogether. Billy never misses, so that if Suzy's rock had missed, the bottle would certainly have shattered. Billy's throw raises the probability of the shattering, but does not cause it. Therefore probability-raising is not sufficient for causation either.

It is not clear whether late pre-emption is fatal for probability theories. The reason this isn't clear is that it isn't clear that no revision of the probability raising view will make the problem go away. What is needed is a revision to the theory that entails that Suzy's rock is, while Billy's rock is not, the cause of the shattering. I will mention two such revisions, viz. the view that events are *fragile*, and the revision to this latter approach – the theory of 'causation as influence' – endorsed by Lewis.[185] The fragility approach holds that events are not modally robust, in the sense that they could not occur in a different manner to their actual manner of occurrence without being different events. According to this approach, Suzy's throw, and not Billy's, is the cause, because had Billy's caused the bottle to shatter, it would have been a *different shattering*. Thought of in this way, Billy's throw does not raise the probability of the particular shattering caused by Suzy's rock to anywhere near the degree that Suzy's throw does, if indeed it raises the probability of this shattering at all.[186] There are two central difficulties with this approach. First, the view that events are fragile prevents makes a lot of our ordinary discourse about events come out false. We often talk of *particular events* being delayed, or altered, for instance, rather than of *different* events that might have occurred in place of the actual ones. More

[185] See Lewis [2000] pp.185-9 for a discussion of the merits and drawbacks of fragility and the analysis of causation in terms of influence.

[186] I see no reason to assume that it is *impossible* for Billy's rock to cause the bottle to shatter in exactly the same manner as Suzy's, unless events have their causal histories, as well as their intrinsic properties, essentially. Either way, it will be extremely *improbable* that the shattering will have the same intrinsic properties if caused by Billy's rock rather than Suzy's.

seriously, fragility of the effect makes us count just about everything that occurs prior to it among its causes. The gravitational effect of Billy's rock on the motion of the shards of glass produced by the impact of Suzy's rock, for instance, will alter the manner of the shattering. If it had been raining at the time of impact, then the shards would have been wet, so the rain would have counted as a cause of the shattering. And so on. As a result of these problems, Lewis proposes the following revision to the fragility approach. Rather than treat events as fragile, he treats causation between robust events in terms of probability relation between fragile *alterations* of those events.[187] An alteration of an event is taken to be a difference in the time and/or manner of its occurrence. Lewis thinks of causation in terms of *influence*, and says that:

> C *influences* E if and only if there is a substantial range $C_1$, $C_2$…of not-too-distant alterations of C (including the actual alteration of C) and there is a range $E_1$, $E_2$…of not-too-distant alterations of E, at least some of which differ, such that if $C_1$ had occurred, $E_1$ would have occurred, and if $C_2$ had occurred, $E_2$ would have occurred, and so on.[188]

Lewis's thought seems to be that if the manner of Suzy's throw is varied (e.g. by using a heavier rock), then the manner of occurrence of the shattering varies correspondingly; on the other hand, varying the manner of Billy's throw would make no difference to which alteration of the shattering occurs. So Suzy's throw does, while Billy's throw does not, influence the shattering. I am prepared to accept this, but only on the proviso that the *times* of Suzy's and Billy's throws are invariant. The reason is simple: if temporal alterations are allowed, then Billy's throw will influence the shattering as well. All you have to do is make it ever so slightly *earlier*. But if this is allowed, then there will be a range of alterations of Billy's throw (all occurring at or before a given time very close to the actual time of occurrence) that are related to a range of alterations of the shattering. But now it looks as though influence is far too cheap to do justice to the intuition that Billy's throw does not cause the shattering. The trouble for Lewis's theory, of course, is that prohibiting temporal alteration looks

[187] Lewis [2000] pp.189-91.
[188] Lewis [2000] p.190.

terribly ad hoc. Why allow the 'whether' and 'how' of causes to be relevant to the influencing relation, but not the 'when'? The influence theory hands out *far too much* influence to be able to distinguish pre-empting from pre-empted causes.

The intuitively obvious answer to these problems is, of course, the process account. Billy's throw fails to qualify as a cause of the shattering because process-linkage is necessary for cassation, and the process that would link Billy's throw to the shattering is cut off by the impact of Suzy's rock. No conserved quantity is transferred from Billy's rock to the bottle; however a causal process links Suzy supplying energy to her rock with the bottle's demise. Process accounts get pre-emption cases right, for they allow us to distinguish pre-empted from pre-empting cause. As I said, it is unclear whether probability accounts can be modified in order o accommodate the problem cases or not; but why bother, when the process account is to hand? Unfortunately, process accounts are faced with problems as well, which – or so I am prepared to argue – are much more serious than the problems for probability accounts. The central problem I want to draw attention to here is what Schaffer terms *causation by disconnection*.[189] Schaffer gives many examples of disconnection, all of which, as Schaffer admits, have a common structure. Causation by disconnection is a species of *double-prevention*, which is causation of an effect by preventing something that would have prevented the effect. Disconnection cases are specially designed to cause problems for the process account, for causing by disconnection involves cutting off a *process* that would prevent the effect if left to run to completion. As all the examples Schaffer gives have this structure, I will focus on just one.

A ship is sailing into bad weather, and will almost certainly sink unless a radio message from the local weather centre, warning the captain to turn back, gets through. The captain's wife, fed up with being married to a ship's captain, decides to cash in on her husband's life insurance policy, and sabotages the radio transmitter. This is a case of causation by double prevention – the captain's wife causes the ship to sink by

[189] See Schaffer [2000] for detailed discussion.

preventing the arrival of a message that would have prevented the sinking. However, there is no process linking her sabotage to the sinking of the ship. In fact, this kind of causation is characterised by the very *absence* of such a process – it relies on the fact that no information gets through to the ship. Intuition is clear that the captain's wife causes her husband's death, but she does so by disconnecting a process – there is no process-linkage connecting her action to the sinking of the ship. There is no option here other than for process theorists to deny that the captain's wife causes the ship to sink. But that is a hugely implausible denial; she is clearly morally responsible for the deaths of the crewmen, for she acted knowing that her action would cause the ship to sink. Further, had she not acted as she did, the ship would not have sunk.

As Schaffer notes, causation by disconnection is not confined to thought-experiments; rather, it goes on all the time. For instance, people sometimes, sadly, die *because* their hearts stop. Heart attacks cause the death of the brain and other organs not by sending stop signals to them, but by interrupting the process that supplies these organs with oxygen. The breaking of a levee causes a flood by cutting off a process by which the water was prevented from flowing. You fire a gun (if you are that way inclined) by causing a catch to release the trigger.[190] Process accounts give us *spectacularly* wrong results in such cases, for they must deny that the disconnecting causes *are* causes.[191] If process-linkage is necessary for causation, then we must find another way of talking about disconnection, for disconnecting causes are evidently not process-linked to the distal events that causal intuition clearly regards as their effects.[192] Here, of course, is where probability accounts step in. Although there is no process linking the

[190] See Schaffer [2000] for persuasive arguments that disconnection is ubiquitous.

[191] I note in passing that cases of causation by disconnection are not the only kind of intuitively clear cases of causation that process accounts (at least those that maintain that there is a flow of causal work from cause to effect) must deny. For instance, suppose you cause your head to cool down by placing a block of ice on it. Intuition recognises two causal relations here – the block cooling your head, and your head heating the block. But process accounts must deny that the former of these is genuinely causal, as the flow in that case would be from effect to cause. I have not focussed on such cases because they strike me as far less problematic for process accounts than cases of causation by disconnection.

[192] If a counterexample to the sufficiency of process-linkage for causation is wanted, then we have cases of what Schaffer terms *mis*connection, where a process links events that are not causally related. For instance, it's raining when Suzy throws her rock, and the rock gets wet on the way to the bottle. The rock's being wet is process-linked to the shattering, but is not a cause of it. See Schaffer [2001] for full discussion.

actions of the captain's wife to the sinking of the ship, her action is a probability-raiser of the sinking, and qualifies as its cause that way.

Thus far we have considered counterexamples to the claim that probability raising is necessary for causation, and counterexamples to the claim that a causal process is necessary for causation. Why is any of this important to the causal exclusion argument? The exclusion argument, as I argued in 5.2 and 5.3, depends on process-linkage being a necessary condition for causation. Without this condition, for instance, the exclusion principle $O_D^4$ defined n 5.2 cannot be derived. And the exclusion argument of 5.3 does not go through, for it is only if the causation of M* by M requires process-linkage that the sufficient physical process from P to P* excludes M's efficacy. If causal work in understood in terms of physical work – and it is difficult to see how else to understand it – then "doing the causal work" required or an effect simply *cannot* be necessary for causing it, on pain of having to deny that a great many *apparently* (I am tempted to say self-evidently) causal relations *are* causal. Now *unless* we understand causal work in terms of something like the transmission of conserved quantities, then it isn't clear that there is any causal exclusion problem – there's no reason why the 'causal work' of raising the probability of an effect should not be done many times over. For my part, I find this a compelling reason for thinking of causal work in terms of physical work. But understood in *this* way, disconnection cases give us clear counterexamples to $P_L$ – there are sufficient causes (like the sabotage by the captain's wife) of certain effects (like the sinking of the ship) that don't do *any* of the causal work necessary for the occurrence of those effects. What causal work the wife does consists in preventing the radio signal from preventing the rocks from doing *their* causal work, of making a big hole in the ship. Notice that where there is causal work to be done, there is *something* that does it, in this case the rocks that the ship bumps into. But that does not alter the fact that there is a cause of the sinking that does *no causal work at all* on the ship.

I anticipate an objection at this stage. A proponent of $P_L$ might consider weakening it so that rather than claiming that causes must be process-linked to their effects, $P_L$

claims that in order to cause *anything*, a cause must be process-linked to *something*. Even disconnecting causes, after all, do *some* causal work, even if this is not the causal work necessary for the occurrence of their effects. This weakened version of $P_L$ accepts that the captain's wife causes the ship to sink, and holds that she does this by doing causal work on the radio transmitter rather than the ship itself. Now this sketch of a theory has the promise of once again giving rise to the exclusion problem, for as we saw in 5.3, given the sufficient causal process from P to P*, there isn't anything left over for M to do any causal work on. I will not go into any great detail as to how the revised theory might work, for it fares no better with disconnection cases than the original. To see this, consider *how it is* that the action of the captain's wife causes the ship to sink. As we saw, her action causes the transmitter to stop working, but it is this event in turn – the transmitter's being broken – that causes the ship to sink. If an argument for this point is needed, then simply reflect on the fact that if the transmitter had malfunctioned of its own accord, this too would have been a sufficient cause of the sinking. But now we have a cause – the broken transmitter – that causes its effect without doing any causal work at all; this is not surprising, for it causes the ship to sink precisely because it *stopped* working! The central rebuttal to the revised process-linkage theory is that disconnection involves, at some stage, causation by the *non*-occurrence of certain events. Disconnection, as I said, is a special cause of causation by double prevention, and as such is bound to involve the non-occurrence of the prevented preventer *as a cause*.[193] In the case of the transmitter, the captain's wife causes it to malfunction and its failure to send out the signal causes the ship to sink. The ship sinks because of something that failed to happen, and something that does not happen can do no causal work. If disconnection cases are counterexamples to the original $P_L$ – and they clearly are – then they will also be counterexamples to the weakened version.

[193] Causation by absences is one of the main reasons why Mellor [1995] endorses the view that the relata of causation are facts rather than immanent particulars such as events or states of affairs. If I am right that double prevention always involves causation by an absence, then there will be a great many cases of causation where one of the relata is *missing*. I am not sure whether this difficulty can be overcome within a Kimian metaphysic of events; my intuition tells me it cannot. These matters are beyond the scope of the present work. For now, note that causation by absences is a counterexample to the claim that causes must do some causal work, for absences by their very nature can not.

Now if these remarks are correct, then $P_L$ is straightforwardly false: process-linkage is not necessary for causation. Of course it may be that $P_L$ is not what Kim and others have in mind when they worry that there is no causal work left for M to do in causing M*, given that P is causally sufficient for P*. Perhaps there is some other way of conceiving causal work such that mental events are not, while disconnecting causes are, able to do it. I am unable to see what this alternative might be. The theory of causal work as process-linkage that I detailed in 5.2 recommends itself as a correct interpretation of those who think there is a causal exclusion problem not least because it enables us to derive the otherwise unsupported strong principle of causal exclusion, $E_X$, which we saw Kim explicitly appealing to in the supervenience argument of 5.1. Further, understanding causal work as physical work has significant scientific respectability. The causal exclusion argument as I understand it gets causal work right, but gets causation wrong. For causal work, construed as physical work, simply isn't necessary for causation. As such, it does not *matter* whether or not there is any causal work left for mental properties to do, for they don't *have* to do any in order to qualify as causes of behaviour. Notice that it seems as though process-linkage accounts are considerably worse off than probability accounts when it comes to accommodating intuitively clear cases of causation. This is because it is not clear that the pre-emption problem for probability accounts cannot be remedied without appealing to process; by contrast, it is extremely difficult to see how to solve the disconnection problem for process accounts without appealing to probability. Causes that have their effects by disconnection do not seem to be connected to their effects by any process at all, so it is unlikely that tweaking the way in which 'causal process' is defined in $P_L$ will enable it to accommodate disconnection. Unless, of course, the definition involves an appeal to a kind of process that is really just probability-raising in disguise.

I do not claim that there is nothing that is right about the process account – quite the contrary, for it is possible that probability accounts cannot be patched up internally, and that the only way to properly distinguish causes from non-causes in certain cases

(for instance pre-emption) is to introduce an element of process into the mix.[194] Many – perhaps even *all* – causal relations *involve* causal work in some way. However, if causal work is understood in terms of physical process, then it is simply false that all sufficient causes do the causal work necessary for their effects. And this is all we require in order to solve the causal exclusion problem. Recall our six possible arguments of 5.3, whereby any five of the following propositions can be endorsed to show that the other is false: (i) $E_M$, (ii) $C_P$, (iii) ¬ID, (iv) $P_L$ (v) CW, (vi) $T_{CW}$. Well, now we see that there are *independent* reasons for rejecting $P_L$, and so given the plausibility of the other five premises, we have a very strong case for against $P_L$. That is, we have an argument for $\neg P_L$ whose premises are independently justified, and a stock of very plausible counterexamples to $P_L$ that are not dependent on endorsing any of the premises of that argument.

Before proceeding, I will pause to clear up an understandable confusion, based on the arguments I have given in chapters 4 and 5: "haven't you been endorsing Yablo's proportionality theory, and other probability accounts, up to this point? But now you have given reasons *not only* for rejecting process accounts of causation, but *also* probability accounts such as Yablo's". I agree that the pre-emption cases discussed above provide good reasons to reject the necessity of probability-raising for causation. However, nothing in the causal argument of 3.4 depends on *any* particular theory of causation – we can now see just how considerable a virtue this is. And when I appealed to probability accounts in chapter 4, I did so merely to show that transmission principles *are* dependent on theories of causation in a way that the causal argument should *not* be. Similarly, my appeal to probability accounts in 5.4 was intended merely to show how endorsing such accounts (and hence denying $P_L$) enables us to avoid the exclusion problem. Notice that the conjunction of the five theses that I endorse as premises against $P_L$ is perfectly consistent with the *falsity* of

[194] Schaffer [2001] takes this to suggest that a hybrid account is needed, and according to the account he gives, causes raise the probabilities of processes connected to their effects. This is not the place for a discussion of this interesting theory. Note that the manner in which process and probability accounts seem to get each others' problem cases right may well be indicative of an inherent inhomogeneity in the very concept of causation. Perhaps the concept is multifarious, so that causal relations sometimes involve process, sometimes probability, and maybe sometimes *neither*. See Hall [2001] for discussion.

probability accounts. True, two of the other premises concern causal work (CW and $T_{CW}$), but they make no claims about the relationship between causal work and *causation*. CW claims that the work is only done once, which it is difficult to deny; $T_{CW}$ says that the work necessary for a dependent effect is the same as the work necessary for the effect it depends on, which is also difficult to deny. Rejecting $P_L$ therefore, neither requires nor entails any *other* particular theory concerning the nature of causation.

The causal exclusion argument, if it were sound, would provide us with an argument for type identities; we could endorse $E_M$, $C_P$, $P_L$, CW, and $T_{CW}$ in order to prove ID. But $P_L$ is false, and so this argument isn't sound. In conclusion to the present chapter, then, we are back where we were at the end of chapter 3: the causal argument proper (based on a general principle of non-overdetermination not parasitic on any theory of causation, let alone one as dubious as $P_L$) establishes that physical events are synchronically sufficient for mental events. This in turn licenses a form of supervenience that will be physicalist provided the sufficiency between the events holds across all physically possible worlds. In the next chapter, we will see that there are weaker forms of sufficiency consistent with the premises of the argument, and that as a result, the argument is invalid.

# 6. Emergence, Novelty and Redundancy

My purpose here is to give a general characterisation of emergence, and show that there is room for a weaker version of emergence than is usually acknowledged. I diverge slightly from received wisdom in that my conception of emergence avoids epistemological notions like reduction and prediction. It should be noted that my purpose here is not to defend any form of emergence; rather, I aim only to defend its possibility. I do not claim that the form of emergence I describe is endorsed by any of the main proponents of emergentism, although it seems to me to encapsulate many of their central ideas. As I will conceive it, emergence is the conjunction of a metaphysical claim concerning the nature of the relationship between physical and emergent properties, and a claim about the novel causal powers introduced by emergent properties. In 6.1, I will outline the metaphysical commitments that I take to distinguish emergence from physicalism. In 6.2, I argue for a distinction between two kinds of novelty, and show how this distinction enables us to resist redundancy arguments such as the one due to Kim, which I introduced in 3.3. In 6.3, I explain how, on the basis of the theories of 6.1 and 6.2, three kinds of emergence can be distinguished, one of which – 'weak' emergence – is consistent with the premises of the causal argument. I conclude on this basis that the causal argument is not valid, and that further arguments must be supplied if the argument is to establish physicalism. I suggest two such arguments, which, when taken together, provide a compelling case against weakly emergent mental properties: an epistemic argument in 6.4, and a teleological argument in 6.5.

## 6.1. Metaphysics of emergence

I take Broad's [1925] brand of emergentism to be fairly close to what I have in mind, but nothing of import for my purposes turns on whether or not this is so. Provided the position I outline is consistent, then the central arguments of chapters 6 and 7 will go through regardless of how similar my emergence is to anyone else's. Emergentism conceived as a metaphysic of mind has much in common with physicalism. It is mental *properties* that are emergent; there is no emergent 'mental substance'.

Emergentism thus shares with physicalism the thesis of physical monism. Further, emergentism also shares with physicalism the view that the emergent mental properties are determined by, and supervenient upon, physical properties. The central metaphysical distinction between physicalism and emergentism is that the former affirms, whereas the latter denies, that mental properties are nothing over and above the physical. I follow Broad in supposing that if there are emergent properties, then they are properties of aggregates of physical particulars, synchronically determined according to what Broad terms 'trans-ordinal laws' that take the structural properties of these aggregates in the antecedents and have emergent properties in their consequents.[195] The laws in question hold independently of the laws of physics – they are 'unique and ultimate' laws true in some physically possible worlds, not in others.

Now the supervenience of emergent properties on the physical means that great care must be taken to distinguish emergentism from physicalism. Because emergent properties and properties for which physicalism is true both supervene (as we shall see, in very similar ways), it is quite common in the literature to find emergentism defined by way of epistemological claims, for instance that emergent properties cannot be predicted from, or functionally reduced to, physical properties.[196] However, there is a metaphysical difference that makes all the difference – trans-ordinal laws are not physically necessary. From this it follows that if there are emergent properties, then they will not be instantiated at minimal physical duplicates of the actual world. Similarly, although physical properties will be sufficient for the emergent properties, the sufficiency relation will be nomologically, but not physically, necessary. And this is precisely what makes emergent properties something over and above the physical. For my part, I take this modal difference to be the defining metaphysical characteristic

[195] See Broad [1925] pp.77-80 for discussion of trans-ordinal laws.
[196] See for instance Stephan [1997], Beckermann [1992] and Kim [1999a] for endorsement of the view that the metaphysical component of emergentism is to be interpreted epistemologically. All three essentially hold that what distinguishes emergence from physicalism is (*inter alia* in Kim's case, as he combines the epistemic claim with a claim about *downwards causation* – of which more presently) that physical properties are, while emergent properties are not, functionally reducible to the physical. Stephan talks in terms of superdupervenience rather than functional reduction, but as I argued in 2.1, much of what Horgan has to say about superdupervenience suggests that functional reducibility is what he has in mind.

of emergence; the irreducibility of a property is a *necessary, but not sufficient*, condition for its emergence.

By way of illustration, consider once again Kim's strong supervenience of the Ms on the Ps:

$$\Box\forall x\forall \mathrm{M}\in M\{\mathrm{M}(x)\rightarrow\exists \mathrm{P}\in P[\mathrm{P}(x)\&\Box\forall y(\mathrm{P}(y)\rightarrow\mathrm{M}(y))]\}$$

If mental properties are emergent in the sense described above, then they will satisfy strong supervenience *provided the strength of the second operator is no greater than nomological*. This is precisely why the second operator must express at least physical necessity if the definition is to be considered as a definition of physicalism. For as I have argued, nothing forces the view that all laws of nature are, or are determined by, laws of physics. And if Broad is right that there are emergent properties, then some laws of nature are neither laws of physics, nor physically necessary. A law fails to be physically necessary just in case there are physically possible worlds at which the law does not hold. Alternatively (and equivalently) if there are laws that are not physically necessary, then minimal physical duplicates of the actual world will be worlds at which some of the actual laws of nature do not hold. Failure to properly distinguish the different supervenience theses resulting from alternative modalities can, I think, lead to confusion regarding the difference between non-reductive, supervenience physicalism, and emergence. Crane, for instance, argues that strong supervenience is unable to distinguish emergentism from physicalism, by arguing that both satisfy strong supervenience. This claim is true as far as it goes, but it seems clear to me that it does not go far enough: emergentism and physicalism differ as to the strength of the second modal operator. How does Crane interpret this operator? He doesn't say. Consider the following passage:

> The notion of supervenience [of a set A of properties on a set B of properties] does not say anything about whether the A-properties are "something over and above" the B-properties: [strong supervenience] is consistent with the distinctness of the of the A- and B-properties, and also consistent with the identification of each A-property with a B-property. In addition, it is

> consistent with the A-properties having independent causal powers. So, the strong supervenience of the mental on the physical is consistent with emergentism.[197]

I agree that the mere *notion* of strong supervenience is consistent with emergence, and also with the B-properties having independent causal powers, but I think Crane gives far too little importance to the interpretation of the modal operators: whether or not the mental is something over and above the physical is determined (or so I argued in chapter 1) by the strength of the second (sufficiency) operator. If it is interpreted as physical necessity, then physical properties alone, together with the laws of physics, will be sufficient for the mental properties. But that means that minimal physical duplicates of the actual world preserve the actual distribution of mental properties – and if this much is true then, as I have argued, there is a clear sense in which the mental is nothing over and above the physical. If, however, the strength is merely *nomological*, and if, in addition, the set of natural laws contain laws not determined by physical laws and properties, *then* the definition is consistent with emergence, and with the mental being something 'over and above' the physical. If the mental emerges in this way, then minimal physical duplicates of the actual world are not duplicates *simpliciter*, as the extra 'trans-ordinal laws' by definition will not obtain at minimal physical duplicate worlds. Thus I hold that the distinction between nomologically and physically necessary supervenience conditionals is a difference that makes a crucial difference when it comes to distinguishing physicalist from non-physicalist positions. Crane's purpose is to argue that (i) if non-reductive physicalism and emergentism share all the same metaphysical commitments, any problems for emergentism consequent upon its particular metaphysical commitments must also be problems for non-reductive physicalism, and (ii) they *do* share all the same metaphysical commitments.[198] It is difficult to deny (i); for my part, however, I find it equally

[197] In his [2001].

[198] In particular, Crane [2001] argues that both non-reductive physicalism and emergentism hold that mental properties are supervenient upon, and distinct from, physical properties. Further, he agrees with Kim [1992b] that the only way to give content to the distinctness part is via the claim that the supervenient properties are novel, which in turn means they exert a downwards causal influence on the physical domain. This is the crux of Kim's dilemma for supervenient causation: either the supervenient properties are redundant, or they violate the completeness of physics. I will consider and reject this argument in 6.2. I should note in passing that while I do not agree with Crane that emergentism and

difficult to endorse (ii), as supervenience formulations of non-reductive physicalism need physically necessary sufficiency at least, on pain of not being formulations of *physicalism* at all.[199]

Beckermann [1992] holds similar views to Crane on this matter, which I reckon is also due to a failure to properly interpret the relevant modal operator. Beckermann too argues that strong supervenience does not discriminate between emergentism and non-reductive physicalism. For Beckermann, though, emergentism is consistent with strong supervenience, even if the strength of necessity in the second operator is taken to be *physical*.[200] If emergentism is consistent with strong supervenience so interpreted, then by my reckoning, emergent properties are not emergent at all. It seems that Beckermann is tacitly assuming that physical and nomological necessity are the same – but this, as I understand it, is one of the claims emergentism denies. Noordhof's position on these matters is very close to my own.[201] He argues that physicalists employing strong supervenience to define their position must interpret the strength of the second operator as *metaphysical*, "otherwise there really would be no way of distinguishing materialism from British emergentism". But for Noordhof, metaphysical necessity is not importantly different from physical necessity as I have conceived it. This is because Noordhof allows that laws of physics into the supervenience base. But the claim that a physical property P *together with the laws of physics* metaphysically necessitates a mental property M just means 'it is metaphysically necessary that if the laws of physics hold, then if anything is P then it is Q'. But this can be re-written as 'in every metaphysically possible world, it is true that in every physically possible world that if anything is P then it is Q.' The wide-scope quantification seems otiose here: why not just appeal to *physical necessity*

---

non-reductive physicalism share the same metaphysical commitments, I do not think that the metaphysical distinctions I have drawn need trouble Crane's overall argument. The crucial similarities between emergentism and non-reductive physicalism, for Crane's purposes, are a common commitment to (i) distinctness, and (ii) novelty, of mental properties. These commitments, I agree, are common to both positions, despite their metaphysical differences.

[199] In 6.4 I will show that the different modalities of emergence and supervenience physicalism leaves a version of the former (weak emergence, to be defined in 6.3) open to a redundancy argument that does not affect the latter.

[200] Beckermann [1992] p.103, fn.11.

[201] See Noordhof [2003] pp.85-93.

instead? This matter is terminological. What is important here is that Noordhof, like me, thinks that depending on the strength of the second modal operator, strong supervenience can define either physicalism or emergentism. But no fully interpreted strong supervenience thesis will define *both*.

It remains for me to comment on the relationship between functional reducibility and emergence. Beckermann defines emergentism like this:

> Let *S* be a system having the microstructure [$C_1$,…,$C_n$; R], then *F* is an emergent property of *S* iff (a) there is a law to the effect that all systems with this microstructure have *F*, but (b) *F* cannot, even in theory, be deduced from the basic properties of the components $C_1$,…,$C_n$ and a general theory of components of this kind which contains no unique and ultimate laws which apply only to systems which have the same microstructure as *S*.[202]

For Beckermann, deducing a property is very similar to functionally reducing it, and for present purposes I will take it that functional reduction is what he has in mind – we first construe *F* in terms of its causes and effects, and show how $C_1$,…,$C_n$ plays the role of *F* by reference to the laws that govern the components of *S*, regardless of whether or not they are combined according to the 'system-defining' relation R. The stipulation that the laws of the realizer theory contain no unique and ultimate laws that apply only to R-systems serves to rule out the trans-ordinal emergence laws from counting among the premises from which *F* is deduced. Beckermann's thought is that *without* these special laws, there is no way to deduce an emergent property. I am in agreement that (a) and (b) above are *necessary* conditions for emergence. This much ought to be clear from the account of functional reduction I gave in chapter 2 – if *F* can be functionally reduced to basic (let's say physical) properties and laws, then *F* is nothing over and above the physical. However, I am unable to agree that (a) and (b) are sufficient for emergence; I will briefly explain why.

[202] Beckermann [1992] p.106. This definition says nothing about the causal novelty of emergents, and in fact Beckermann has nothing to say about this element of emergentism. It may be simply that he takes the novelty of a property to be built into its very existence condition: a property that has no novelty is no property at all. We return to this matter in 6.2 when we discuss redundancy arguments.

The central reason why I wish to resist the equivalence of emergence and non-deducibility proposed by Beckermann is that it is possible that physicalism is true for all actual world properties (so that they are nothing over and above the physical, in the sense that all minimal physical duplicates of the actual world are duplicates *simpliciter*), and yet certain properties not be functionally reducible to the physical. The point I wish to make here has nothing to do with the difficulty or length of the relevant deductions. As soon as any such appeal is made, those who like to do metaphysics with epistemological notions will appeal to 'deducibility-in-principle', which is the sort of thing a superbeing could pull off *in practice*, given enough time and coffee. Rather, I wish to point out that some properties, by their very natures, may well *fail* to be functionally reducible. Suppose physicalism is true. Given physicalism, then on reasonable further assumption, it is plausible that properties for which step (1) – *functionalization* – of the functional reduction process can be completed, *can* be deduced.[203] A functionally defined property F that has a physical realizer P will be deducible from P, provided it follows from physical laws that P plays the appropriate causal role R. I am prepared to grant that this point for the sake of argument – let's agree that *given physicalism*, the causal roles of *all* physical properties are determined by physical laws, and as a result it will be possible to deduce F from P for any functionalizable F.[204]

The crucial thing to realize, however, is this: there is no law of physics that states that all properties are functionalizable. The assumption that any property that is nothing but the physical, is also a property that can be accurately reconstrued as a second order functional property, is substantive and to my mind wholly unjustified. Far

[203] See 2.2 for discussion of the functional reduction process.

[204] I stress 'given physicalism' for two reasons. *First*, it might be argued that certain kinds of emergent property confer causal powers in addition to those of their base properties that enable these latter to play causal roles that go beyond those determined by the physical laws in which they figure. (Wilson's argument of 1.3 depends on just this sort of view.) Myself, I do not agree – as I will argue in 7.3, this situation is best described as one in which emergence base and emergent property *together* play the role in question. *Second*, I think it plausible that some emergent properties *are* functionalizable, at least in the sense of its being possible to reconstrue them as causally individuated. Such a property will fail to be functionally *reducible*, however, on the grounds of not being physically realized – the reduction will fail at step 3. See section 2.2 for details of the steps involved. I will comment briefly on the second issue below.

better, in my opinion, is for a metaphysical definition of emergence to leave open the possibility that there are properties for which physicalism is *true* and yet about which we could not (even in principle) *establish* physicalism by reduction. Despite my allergy to epistemic metaphysics, I do of course think that metaphysical notions are intimately related to epistemic ones. Here are some propositions that partially characterise the relationship in the case of physicalism, emergence and functional reduction: (1) a property reducible to the physical is nothing over and above the physical; (2) it is possible for there to be properties that are not reducible to the physical but which are, nonetheless, nothing over and above the physical; (3) if a property is emergent, it is not reducible to the physical.

Before proceeding, I want to elaborate a little on proposition (3) above. Irreducibility is uncontroversially necessary for emergence, but where does the reduction *fail* in the case of, say, emergent *beliefs*? Psychological properties are plausibly (at least partially) individuated by their causes and effects; as interpreters, we rely on this fact to facilitate belief ascription. But isn't that all we need in order for functional reconstrual to get off the ground? What room does that leave for emergent psychology? The answer lies in recognising that there is a crucial difference between the following two claims (i) F is individuated by causal role R; and (ii) F is the property of having some property that plays causal role R. We cannot infer (ii) from (i), as there may well be properties that have their causal roles essentially, but that are not realized by any *other* properties. Basic physical properties such as charge might be a case in point: for instance, charge is arguably partially individuated by its causal role as specified by physical laws. But *qua* (putatively) basic, then the 'charge role' is played by *charge*, and not some distinct role-filler property. Of course if charge has its causal role essentially, then we can reconstrue it as a second-order functional property. But such a reconstrual will be incorrect, as there is, by hypothesis, no lower-order realizer of charge.

What then of causally individuated but emergent psychological properties? I think it clear that we can complete step (1) of the reduction process for such properties. There

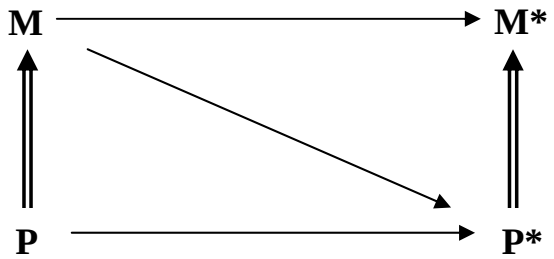
is nothing in the nature of our hypothesised emergent mental properties that prevents their being (incorrectly) reconstrued as second-order functional properties. Step (2) can also be completed, for the emergence bases of our emergent mental properties will provide excellent putative realizers. Now it follows that for causally individuated emergent properties, step (3) must fail, on pain of the properties in question not being emergent at all. There are two reasons why step (3) might fail for emergent properties, corresponding to my distinction (to follow in 6.3) between weak and strong emergence. In the former case, step (3) fails because the individuative causal role R is specified in terms of *other* emergent properties; in the latter case, it fails because R involves physical effects that the putative physical realizers do not have. Either way, step (3) fails because the putative realizers do not play causal role R, so that F cannot be deduced from P. I will clarify and defend this position during the course of the next two sections.

The central point I hope to have made in the preceding paragraphs is that emergence can be distinguished from physicalism at a purely ontological level, without recourse to epistemological notions. It is very much like physicalism, except that emergent properties supervene on physical properties *and trans-ordinal laws*, which latter are not determined by the laws of physics. Emergent properties, though, are also *novel*, in a sense I have been promising to define. Let us turn, then, to the crucial issue of the sense in which emergent properties are novel. Before proceeding with this task, I will describe two ways in which properties in general can be novel. Through this, I will explain under what conditions I take a property to be redundant, and so explain why the redundancy argument suggested in 3.3 is easily resisted. Following brief consideration of novelty and redundancy in general, we return to the issue of the sense(s) in which emergent properties are supposed to be novel.

## 6.2. Two kinds of novelty – why the redundancy argument fails

My aims in this section are (i) to convince you that there are (at least) two ways for a property to be novel (i.e. non-redundant); and (ii) to show how recognising two kinds of novelty enables us to resist the redundancy argument suggested in 3.3. Before

proceeding, we need to clarify how the redundancy argument works. The redundancy arguments to follow will be directed against all supervenient properties, on the assumption that the completeness of physics is true. I will first give a redundancy argument that combines two elements of Kim's thinking on these matters.[205] I will then give an improved argument, and show how it can be resisted. The first argument places general constraints on what it is for a property to be novel, then proceeds to show that given $C_P$, supervenient properties are unable to meet those constraints. The following diagram will suffice to illustrate both Kim's version and mine:




The diagram is no doubt familiar by now. The fundamental things apply: M supervenes on P, M* on P*; M and M* are instances of novel properties, non-identical to P and P* respectively; M causes M*, and P causes P*. It should be noted that nothing in what follows depends on the relationship between M and P being supervenience (although I will speak of P as M's base property) – what is important is that M≠P. Further, nothing in the argument depends on the supervenience of M* on P* being of any particular strength – as we shall see, it is consistent with M* being an emergent property. Now, 'redundant' – at least as I understand it – means 'not required'. We will be concerned with causal (and causal-explanatory) redundancy. To say that something is causally (or causal-explanatorily) redundant is to say that it isn't required to cause (or causally explain) anything. Which, in turn, is to say that its

[205] It will be clear to anyone familiar with Kim's work that the premises upon which the argument depends are endorsed at various places throughout his work; equally clear, I think, is that Kim would endorse the argument that I give. Still, I should point out that I am not aware of any explicit presentation by Kim of the argument to follow, in its entirety, in exactly the same form as that in which it occurs here. The elements I combine are Kim's claims concerning novelty in his [1992b] pp.134-7, and a version of the causal exclusion argument we discussed in 5.1, to be found, *inter alia*, in Kim [1993b] p.354, [1998] pp.44-5.

putative effect (or *explanandum*) *already has* a cause (or *explanans*). The conclusion of a redundancy argument is that ontological commitment to the redundant entity buys us nothing, and that as a result, we should not be so committed. This much is of course uncontroversial; if mental property instances aren't required as causes of anything, then why believe in mental properties? The controversy surrounds whether or not there is any purpose for which mental properties *are* required. Bearing these points in mind, we proceed to argue that M is causally redundant.

Kim's problem be thought of in the following way: how are we to reconcile the supposed novelty of M *qua* non-identical to P with the supervenience of M* on P*?[206] The argument is a simple one, and proceeds as follows. First, Kim endorses a principle he terms 'Alexander's Dictum', which he expresses thus: *to be real is to have causal powers*. I need not take issue with this claim here, for Kim only uses it to conclude that realism about the mental entails that mental properties have causal powers, and I do not dispute the latter proposition. Second, Kim claims that if M is non-identical (in Kim's terminology *irreducible*) to P, then M must have causal powers that are non-identical (irreducible) to the causal powers of P. As we saw in chapter 5, Kim thinks of causal powers in terms of causal *work*; correspondingly, he goes on to say that if M is novel, we must "find for it causal work *not done by the physical and biological properties*" it supervenes upon.[207] This approach is not mandatory, though; I will say that *whatever* properties do, and *however* they do it, if M≠P, then M must do something that P does not. For instance, if we think of the causal powers of a property in terms of the typical effects of its instances, then M≠P demands that M causes something that P does not; or if causal explanatory relevance is paramount, then M≠P demands that M causally explains something that P does not. The general point is that if M≠P, then instances of M must cause (or causally explain) things that instances of P do not. This is a crucial claim, and justification is easier if

[206] Kim states the problem not in terms of non-identity, but in terms of *irreducibility*. For Kim, these two amount to the same thing, for as we saw in 2.3, Kim takes functional reduction to yield *identity*. Since I take functional reductions to yield supervenience, I substitute non-identity for irreducibility in my exegesis.
[207] Kim [1992b] p.135.

we look to its contrapositive: if a property P has all the same causal powers as Q, then P=Q. We need not endorse a Shoemakerian metaphysic of properties in order to support this latter claim; for present purposes, we can be content with epistemological grounds. In particular, note that we could never have any epistemological grounds to *doubt* it – no detector will tell two properties apart if they have all their causal powers in common.[208] What could possibly justify the view that 'they' are two, rather than *one*?

We can now cause problems for supervenient properties, given $C_P$. Consider first the following argument, which we may reasonably attribute to Kim. M's putatively novel causal contribution is that it causes M*. For present purposes, we may consider M* to be either a mental property-instance, or a behaviour. Now given the downwards transmission argument, the *only way* for M to make its novel contribution is through causing P*. As we saw in 5.1, for Kim the same level causal relationship between M and M* presupposes the downwards causation indicated by the diagonal arrow above. Now assuming that $C_P$ is true, P is causally sufficient for P*. If all this is true, then M's putative causal novelty consists in causing an event P* for which there will always be a physical cause P. This is a violation of our novelty principle above – if every M-instance has just the same causal powers as some P-instance, then why not just identify the instances? As we saw in 2.3, instance identity entails property identity, which in turn – given multiple realization – leads to eliminativism. But if M is redundant, then this is as it should be. The downwards transmission argument and $C_P$, make it look as though mental properties confer whatever causal powers are 'already' conferred by their physical supervenience base properties. Together, downwards transmission and $C_P$ entail that the *causal inheritance principle* (CIP) that we examined in chapter 2 is true for *all* supervenient properties, not just the functionally individuated ones. For whatever M's putatively novel causal role is,

[208] See Armstrong [1978] pp.43-4 for discussion of the association between properties and causal powers. Armstrong endorses claims such as the present one for epistemological reasons, and treats them as methodological principles guiding a theory of universals. I note in passing that the principle stated above is intended to apply only to properties that have causal powers – otherwise it would entail that all abstract properties are identical, an unwelcome result.

given downwards transmission it will involve the power to cause a physical event P*, which, given $C_P$, will be caused by M's base property P. The powers of M-instances *have to be* identical to the powers of P-instances, on pain of the M-instances being unable to cause instances of M*. But in that case, commitment to mental properties adds nothing – M is not *required* as a cause of P*, for P* has a physical cause P.

For my part, I reject both CIP and downwards transmission. I reject CIP because, as I said in 2.4, the causal powers of instances of realized properties are just *not* identical to those of their realizer-instances. Much more reasonable is the supposition that the powers of M are a *subset* of the powers of P, which is why I tentatively endorsed CIP' instead. In 4.3, we saw how certain counterfactual theories of causation entail that downwards transmission fails. Now we can see how the subset theory of realization must deny downwards transmission as well. The reason is simple. If the powers of M are a subset of the powers of P, then they will be a subset too of the powers required to cause P*. M *cannot* cause P*, for put simply, it does not have the power to do so! And of course if this is so, then we require an account of causation that allows *intra*-level causation without *inter*-level causation; we have already seen that such accounts are available, and I will not repeat them here. Rather, I will simply note that reliance on downwards transmission makes the above redundancy argument very easy to resist. However, just as we did for the causal exclusion argument in 5.2, we can recast the redundancy argument so that it does not rely on transmission. In order to do so, we must focus on causal *explanatory* redundancy instead of *causal* redundancy. The new argument will not show that we do not need to invoke M as a *cause* of M*, but rather that we do not need M in order to *explain* M*. Things are much simpler this time around, and we need only one extra premise: *that P explains the occurrence of M**.

I maintain that there is a clear sense in which P* explains M*, and that this is true regardless of whether M* is something over and above P*. If P* is M*'s emergence base, for instance, then P* together with the trans-ordinal law that governs M* emergence, suffice to explain M*. But now P, as P*'s cause, will explain M* as

well.[209] We can demonstrate this on the assumption that subsumption under a law is sufficient for explanation. Given P*'s occurrence, it is a non-causal law that M* is instantiated, and there is a causal law relating P to P*. Given the transitivity of nomic sufficiency, then, it seems clear that there will be a law – which, for terminological consistency, we may call a 'law of inducement' – relating the occurrence of P to the occurrence of M*. It follows that there is a sense in which P explains M*. Nothing forces the view that this sort of explanation is causal, but I think it can plausibly be considered as such. The explanation of M* by P combines a causal relationship with a non-causal one – in essence, we explain why M* occurs by causally explaining its supervenience base, and this strikes me as a perfectly good *causal explanation*. If these remarks are correct, then it follows immediately that there is a sense in which M is explanatorily redundant with respect to M*, for we already have an explanation of M* in the form of P. What is more, $C_P$ and the supervenience of M* on P* together *guarantee* that such an explanation will be available. Notice that we can frame the present problem in terms of causal redundancy as well as explanatory redundancy, thus: how can mental properties have novel causal powers if all their effects have sufficient physical inducers? Why is M required as a cause of M*, given that P induces it? This is not, of course, to say that instances of supervenient properties are caused by the causes of their base properties, for that would be to endorse *upwards* transmission. If we wish to avoid upwards transmission, we can put the point like this: given the causal determination of P* by P, and the synchronic determination of M* by P*, M* simply does not *need* a cause of its own.[210]

This redundancy problem is quite general, for given $C_P$, it follows that any event for which a physical event is non-causally sufficient, can be causally explained by citing the cause of its physical base property. I am prepared to concede on this basis that

[209] Similar arguments are to be found in Sober [1999] pp.548-9.

[210] This is one of the possibilities considered in Kim [2003] in response to the causal exclusion problem; see for instance the diagram on p.159. The price of distinctness for M and P is that there is no causal arrow from M to M*; the relationship between M and M* is "like a series of shadows cast by a moving car," (Kim [1998] p.45). The exclusion argument concludes that there *could* not be an arrow from M to M*; the redundancy argument concludes that there *need* not be one. Of course, once this much is admitted, then given Alexander's Dictum, we lose mental realism – inefficacious mental properties, we can do without.

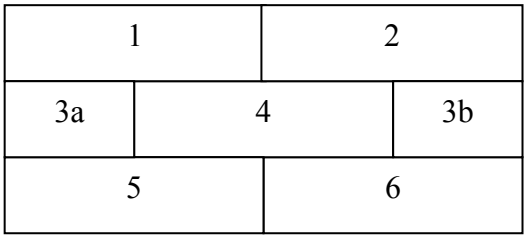
there is a sense in which we do not need M in order to explain M*. What I am not prepared to concede is that there is *no* sense in which we *do* need M. The remainder of this section will argue that while M is not needed in order to explain M*, it is needed in order to explain M* *in a certain way*. Causally explaining M* by citing P, and explaining M* by citing M, are different kinds of explanation. In the absence of an independent reason to prefer one kind of explanation to the other, we ought to embrace both. In the remainder of this section, I will attempt to convince you that M is worth holding on to. The points I raise in support of M's worth are not new, but they are commonly overlooked, quite interesting, and well worth repeating.

The distinction between singular and general causation is familiar. It is clear that there is a relationship between singular causal statement like "David's drinking wine caused him to be intoxicated" and the corresponding general claim that "Drinking wine causes intoxication". Statements of the first kind express causal relations between token events; statements of the second kind express causal regularities between classes or types of events. However, the nature of the relationship has proved difficult to pin down. There seem to be three possibilities. First, we might maintain that there are two independent species of causation, that require different theories. Second, it might be that general causal claims are just generalisations of true singular statements. Or third, perhaps the truth of a singular causal statement depends on its instantiating a general regularity. I will not attempt to choose between these alternatives, however – provided the distinction between particular causal relations and causal regularities is accepted, my central claims here go through regardless of which, if any, is the fundamental kind. Those claims are: (i) that there are two ways in which properties can be novel, corresponding to singular and general causation, and (ii) a property can be novel in the general sense without being novel in the singular sense. The redundancy argument depends, I will argue, on recognising just one kind of novelty, viz. the singular variety, and showing that M lacks this kind of novelty. The position I will describe agrees that there is a sense in which M is redundant, for it agrees that M lacks singular novelty. However, once *general* novelty is admitted, it is clear that the redundancy argument fails, for M *does* have this kind of novelty. Before

defining singular and general novelty, and explaining why this distinction enables us to resist the redundancy argument, I will first attempt to convince you that the distinction is cogent. To this end, I will tell you a story about Bob the builder.

On his days off, Bob the builder likes to keep in shape, and sometimes he does so by building things. Today he has decided to build a wall, out of the six bricks he had been keeping for just such an occasion. Here is a plan of the wall that Bob has decided to build – call this design plan 'D':

All that is important to Bob is that the wall he builds meets D. In order to build a wall according to D, Bob will first have to cut one of his bricks in half. Once he has done so, he can proceed to stick the remaining bricks together in any one of a large number of ways. Number the bricks from 1-6. The diagram below shows he might stick the bricks together in order to build his wall:




There are six bricks that Bob might choose to cut in half, and two ways to arrange the resulting half bricks. There are five further slots for his remaining bricks to occupy. Thus, with his six bricks, there are 6 x 2 x 5 x 4 x 3 x 2 x 1 = 1440 permutations that will allow him to build his wall. (I will ignore additional permutations made possible by rotating the bricks in three dimensions.) This particular permutation has a unique

identifier, as does every other. This one is number 1 2 3a 4 3b 5 6, but Bob cannot tell the bricks apart, and so is unaware which permutation he is assembling. This is unimportant to Bob, because as I said, all that matters to him is that he manages to assemble *some permutation or other* that meets D. Since this is Bob's day off, he had something of a late night last night, and so isn't feeling his best. In particular, instead of his usual 99% success rate with regard to identifying and picking up bricks, today he is only 50% successful. He has a pounding headache, occasionally blurred vision, and a bad case of the shakes. Assume for the sake of argument that by making slight variations in these factors (e.g. the intensity of his headache, whether or not he is able to see brick 6, at some *t*), we can make it so that Bob builds any one of the 1440 permutations. These variations define a set of 1440 close neighbouring worlds $\{w_i\}$, each one of which contains a distinct permutation $D_i$ of Bob's wall.

The important point for my present purposes is that in each member of $\{w_i\}$, Bob's building the wall has the very same psychological cause. In each of the w*i* there will be a different physical explanation of why at that world Bob builds $D_i$; but in each case his building a wall that meets D will be explained simply by the fact that this is the sort of thing he likes to do on his days off. Let M be Bob's wish to build a wall that meets D, and M* the existence of such a wall; let P be the complete physical cause (including hangover) of the particular permutation Bob actually builds, and P* the existence of this particular permutation. The problem at hand is that P threatens to make M redundant with respect to causally explaining M*. I am prepared to accept that this is so – M is not required to explain the occurrence of M*. However, suppose for the sake of argument that at least some forms of causal explanation involve subsumption under a law. M's constitutive mental property subsumes M under a particular set of explanatory laws relating it to events of type M*, *which is invariant across all the* $\{w_i\}$. This is because by hypothesis, Bob has the same psychological properties at each of the 1440 worlds where he builds his wall. The very same particular psychological explanation of M* holds true in each case. By contrast, consider non-actual world $w_{335}$, where Bob builds wall # 1 5 3a 4 3b 2 6. Here the constitutive properties of P will have to be different, for P* is different. For instance,

the realizers of Bob's mental properties will have psychologically irrelevant effects, such as Bob suffering a twitch at *t* that caused him to pick up brick 5 second, instead of brick 2. The actual world physical explanation of the actual world permutation $D_a$ that Bob builds, will not explain why his counterpart builds $D_{335}$ at $w_{335}$.

This defence of the causal relevance of mental properties appeals to what, following Jackson and Pettit, we may term the 'realizer-invariance' of their causal powers.[211] If another example is needed, we can look once again to thermodynamics. Consider a certain aggregate of molecules at a given temperature T. The aggregate in question, as we saw in 2.3, is one specific way to have a certain average molecular kinetic energy; there are many other ways. In other words, the structural property of the aggregate that defines the particular distribution of velocities across its components, is one of many possible realizers of T. Each of T's realizer properties has the power to cause a specific rise in pressure, say, in some other aggregate. There will be laws relating these specific structural properties to equally specific pressure rises. Each law will relate a specific way of being at a given temperature to a specific way to be a rise in pressure. Temperature, one the other hand, will be sufficient for a rise in pressure *regardless of how either the rise or the temperature are realized*. It strikes me that this is a perfectly good case of novelty of causal role – for nothing else plays it! Another way of putting the point is to say that thermodynamic properties capture generalizations that are missed if we describe an ensemble of molecules in terms of their specific velocities, masses and other physical properties. The thermodynamic level, it would appear, contains certain patterns of activity and dependency between its properties that are not mirrored at the level of molecular physics.[212]

---

[211] See Jackson and Pettit [1992b] for discussion. A similar line is taken by Noordhof in his [1997]. There, he argues that a functional property *introduces a novel causal role* that is not introduced by any particular physical realizer, in virtue of having a power that particular realizers lack by definition: viz. the power to cause its constitutive effect *however it is realized*.

[212] This is not, of course, to say that the patterns are not mirrored by *statistical* mechanics; they are, and this is what enables the reduction of thermodynamics to go through. Statistical mechanical patterns will disappear too, if we move down to the (non-statistical) level of molecular aggregates and specific velocity distributions. Statistical mechanical properties such as mean molecular kinetic energy also (it goes without saying) have realizer invariant causal powers.

Things are not so straightforward, however, for it remains to explain why this kind of novelty is worth having. Consider the following rejoinder on behalf of the redundancy argument. M* is a particular, dated, unrepeatable token event, whose occurrence is explained by P. Any other actual instances of supervenient properties will also, *mutatis mutandis*, be explained by the causes of *their* supervenience bases. M's putatively novel causal role seems to consist solely in providing a *modally* robust explanation of M* (or its counterparts) that holds at a neighbourhood of close worlds in which M (or its counterparts) is differently realized. But what use is *that*? Why should we *care* what goes on at other possible worlds? We can explain everything that happens *here* by reference to physical properties alone, so M is, after all, redundant. This argument is not without force, but there are (at least) two responses. *First*, why regard the explanation of M* provided by M as being *in competition* with that provided by P? As we have seen, they are distinct styles of explanation, involving appeals to distinct laws, licensing different counterfactual claims, and so on. For example, Sober suggests that a *complete* explanation of M* might involve "the macro-story, the micro-story, and an account of how these are connected."[213] Sober makes this point in response to Putnam's celebrated [1975] 'Peg' argument, according to which the *micro*-properties of a peg are redundant in explaining whether or not it fits through a hole – what is important is the *shape* of the peg, a macro-property. The 'micro-story' contains lots of irrelevant detail that is rightly left out by the (explanatory) macro-story. I am in agreement with Sober that this argument fails – that the micro-story explains *too much* does not entail that it explains *nothing at all*.[214] But it does suggest a rejoinder to the redundancy argument, in the form of a challenge: given that M and P offer different *kinds* of explanations of M*, then in the absence of an independent reason for preferring one kind over the other, why should *either* of them be redundant? For each will be required, not in order to explain M* *simpliciter*, but in order to explain it *in a certain way*. M will be required to explain

[213] Sober [1999] p.550. A defence of this view can also be found in Jackson and Pettit [1992].
[214] See Sober [1999] p.547 for this line of response to Putnam. Putnam's views on explanation are of course similar to Yablo's views on causation. However, I think Yablo would agree that what I have called 'sufficient inducers' *in some sense* explain the events they induce, despite containing irrelevant causal detail. What he denies is that sufficient inducement is sufficient for *causation*.

M* in a manner that *unifies* other actual and counterfactual wall-buildings; P will be required in order to explain why individual wall-buildings turn out exactly as they do. And what is required, is not redundant.

Our s*econd* response is similar to the first; we simply note that the extra modal information conveyed by an explaining M* in terms of M rather than P, *is* of use. This is because what goes on at other possible worlds determines the truth of actual world counterfactuals. For instance, it is true that if Bob had drunk whisky rather than vodka last night, he still would have managed to build his wall. It is also true that if he hadn't got drunk at all, he still would have managed to build his wall. Psychological explanation has the advantage of explaining the existence of Bob's wall in such a way as to make it transparent why these counterfactuals are true. Explaining it in microphysical terms has the advantage of explaining why it is that Bob builds *this* permutation of D, rather than *that* one. Both of these explanations are interesting, and each is novel. It remains to define the sense(s) in which this is true. In order to facilitate reference to particular events and their constitutive properties separately, I will adopt a more correct notation, shown in the diagram below:


Now let [$x$,P,$t$] and [$y$,P*,$t^+$] be our token physical events causally related, and synchronically sufficient for token (also causally related) mental events [$X$,M,$t$] and [$Y$,M*,$t^+$] respectively. On the basis of the preceding discussion, I propose the following definitions of singular and general novelty:

Singular novelty: A property F of a cause [$x$,F,$t$] has singular novelty just there is a token effect [$y$,G,$t^+$] whose occurrence can not be causally explained without appeal to [$x$,F,$t$].

General novelty: A property F has general novelty with respect to a property G just in case it is a causal law that F-events cause G-events, and substitution for F in this law with any predicate F' expressing a different property results in a different law, or no law at all.

A few notes before proceeding. For a property F to have singular novelty, an F-instance must be required in order to causally explain another property-instance. For a property F to have *general* novelty, on the other hand, F must be required in order to frame a law. Singular novelty is defined for a property with respect to token events; general novelty is defined with respect to *classes* of events. The connection between general novelty and realizer-invariance is as follows: if F is a supervenient and multiply realized property, then it cannot be a law that F-events cause G-events unless F-instances possess the power to cause G-instances *however the F-instances are realized*.

Now given the above definitions, M does not have singular novelty. The reason is that regardless of whether or not *causation* transmits up (or down, or both) sufficiency relations there is, as we have seen, a good sense in which [$Y$,M*,$t^+$] can be causally explained by [$x$,P,$t$]. For [$x$,P,$t$] causes [$y$,P*,$t^+$] and by hypothesis [$y$,P*,$t^+$] is synchronically sufficient for [$Y$,M*,$t^+$]. If this is correct, then we have defined a sense in which *all* supervenient properties are redundant, given the completeness of physics. And the sense, I maintain, is that they lack singular novelty as defined. However, if it is a *law* that M events cause M* events (we may assume it is), then M *does* have general novelty. Here is why. Suppose for the sake of argument that causation transmits upwards, so that [$x$,P,$t$] *causes* [$Y$,M*,$t^+$]. Now I suppose that if this is true, then it will be a law that P events cause M* events. However, the regularity expressed by this law is clearly not the same as that expressed by 'M events cause M* events'. Many M events will *fail* to be P, and yet succeed in causing M* events despite this.

On the other hand, [$x$,P,$t$] will have both singular *and* general novelty.[215] It has the latter because it is a law that P events cause P* events. And it has the former because the occurrence of [$y$,P*,$t^+$] can not be explained without [$x$,P,$t$]. Citing [$X$,M,$t$] will not explain [$y$,P*,$t^+$], as instances of (say) mental properties do not explain the precise *manner* of occurrence of their effects. Bob's desire to build a wall, as we saw, explains his building *some* wall that meets D, but not his assembling the specific permutation of bricks that he actually assembles. Now under what conditions is a property redundant *simpliciter*? Well, one candidate definition of redundancy is obvious:

A property is redundant iff it has neither singular nor general novelty.

If these remarks are correct, then redundancy arguments are much harder to supply than we thought. It is not sufficient to show that all the putative effects of a given cause are caused or explained by something else. Rather, we will have to show in addition that those effects are explained *in the same way* by something else. Since inducers do not explain the events they induce in the same way as the supervenient causes of the induced events, it follows that $C_P$ does not entail that all supervenient properties are redundant *simpliciter*. We can now proceed to consider the manner in which emergent properties are novel. Emergent properties as standardly construed have singular novelty, which violates $C_P$. However, there is room for a weaker conception of emergent novelty, according to which emergent properties are consistent with $C_P$ in virtue of being generally novel but singularly redundant. This, as we will see in the next section, makes trouble for the causal argument.

### 6.3. Three kinds of emergence

In this section, I will distinguish 'strong' and 'weak' varieties of emergence. The strong variety is characterised by a combination of the metaphysics outlined in 6.1

[215] At least on the assumption that P itself is not a supervenient property whose base property would rob it of singular novelty by causing its supervenience base.

and singular novelty; the weak variety by the same metaphysics, together with general novelty. Strong emergence can be further divided into two subcategories, which is why there are three kinds of emergence. Broad intended emergent properties to have causal powers that exert a *downward influence* on the physical domain. We can think of this in the following way. Suppose an aggregate of physical events with an emergent property cause some other event. The causal contribution of the emergent property is such that had it not emerged and yet (*per nomologically impossible*) the physical properties of the aggregate remained the same, then the aggregate would not have had the same physical effects. This is the kind of thing Broad had in mind when he spoke of 'configurational forces'. Suppose you take an aggregate of physical particles. Certain of their properties will, to borrow C. Lloyd Morgan's phrase, be 'additive'.[216] Their inertial masses, for instance, can be summed to find the inertial mass of the aggregate. We can calculate how much force it will take to accelerate an aggregate by adding up the forces it would take to accelerate its components. If there are configurational forces in Broad's sense, then not *all* the causal powers of certain aggregates are like that. If inertial mass were an emergent property of aggregates (which of course it isn't), then we would get the wrong answer by summing the masses of the components. Call this *strong emergence*.

Some clarification is needed at this point. Strong emergence as I conceive it involves what Kim calls 'diachronic downwards causation'.[217] The idea is that a property emerges synchronically from an aggregate, and displays singular novelty by exerting an influence on the physical constitution of some part of the world at a later time. Refer back once again to the familiar diagram at the beginning of 6.2. The reason M lacks singular novelty, as we saw, is that given $C_P$, there will always be a complete alternative explanation of M* available, in the form of P. For M to display *singular* novelty, it must be the case that there is a token event whose explanation requires M.

[216] See his [1923] pp.2-3.
[217] See Kim [1999a] pp.18-34 for detailed discussion of downwards causation, culminating in the upwards-downwards transmission argument purporting to show that even the putative efficacy of strongly emergent properties is pre-empted by that of their physical base properties. More on this in 7.3.

If the event that M is required in order to explain is P*, then this is already downwards causation. But if M has singular novelty with respect to the occurrence of M*, then given that P* is synchronically sufficient for M*, it follows that *M must be required in order to explain the occurrence of P**. This is clearly a form of the downwards transmission argument, but rather than claiming that same-level causation entails downwards causation, the claim I make here is that same-level causation *together with singular novelty* entails downwards causation. I will speculate, in passing, that one reason why Kim is so quick to endorse the downwards transmission argument is that he does not recognise any kind of novelty *other* than the singular variety. If everything else is, in some way or other, synchronically determined by the physical, then the only way for a non-physical property M to exhibit singular novelty is for there to be a physical event whose occurrence requires M. And this in turn leads to Kim's dilemma: either non-physical properties violate the completeness of physics, or they are redundant. But as we saw in 6.2, recognising general novelty in addition enables us to avoid this dilemma, by allowing us to accept a suitably qualified version of the second horn.

Now from this it follows that if there are any strongly emergent properties, then physics is not causally complete.[218] The novel causal powers of emergent properties consist in the fact the that atoms and molecules of the aggregates having them have causal powers that they would not possess in a world where the trans-ordinal emergence laws do not hold; they exert forces that they do not exert just in virtue of their physical properties. One consequence is that if you were to take a minimal physical "snapshot" at time *t* of a deterministic world that contains strongly emergent properties, then the original and the copy will diverge after *t*. A point that I wish to emphasise, however, is that I do not wish to claim that emergent properties somehow *take over* from the physical properties they emerge from. As I think of these matters, the emergence base properties still contribute causal powers; the (physically)

[218] Because there will be physical events that do not have *complete* sufficient physical causes. Kim's upwards-downwards transmission argument (discussed in 5.1) can be marshalled against this point. The argument, in my view, is a poor one; we will discuss it in 7.3.

unexpected effects are due to *extra* powers conferred by the emergent properties.[219] The most sensible way to think about this, to my mind, is to hold that strongly emergent properties *combine* with their physical base properties to cause certain physical effects, which effects would not occur but for the instantiation of the emergents. Physics fails to be causally complete given strong emergence for now there will be physical effects that have only *partial* physical causes: specifying a *complete*, *sufficient* cause will, in certain cases, involve ineliminable reference to (*sui generis*) non-physical causes.

There are two further possibilities given strong emergence. *First*, it may be that the strongly emergent properties always emerge from aggregates with a particular structural property S. This is arguably what Broad has in mind when he speaks of 'configurational forces'. If it is the case that aggregates with S have causal powers that are not determined by their physical properties, then it is plausibly *S itself* that has an emergent *power*. Suppose for the sake of argument that mental properties are emergent, but that they do not have multiple emergence bases, such that anyone with mental property M has structural property S. Nothing in this case prevents the *identification* of M with S. Mental properties will be identical to 'neurostructural' properties with emergent causal powers that are not determined by the powers of their components according to the basic laws that govern their behaviour. The reason I say that this is the sort of thing Broad had in mind is that he claimed, *inter alia*, that emergence occurs in chemical compounds, which have the power to bond with other such compounds, a power which – so Broad claimed – was not determined by the powers of their physical components.[220] *Second*, there is what we may reasonably call

[219] This seems to be very much how Lowe thinks of these matters as well. See for instance his [1993], where he speculates that emergent mental properties exert a co-ordinating influence on otherwise disparate physical events in the brain. It is the emergent mental properties together with their neural emergence bases that cause behaviour. I will have more to say about this in chapter 7.

[220] We now know that this is not so – the chemical bond can be explained in quantum mechanical terms, and so is not emergent. McLaughlin [1992] attributes the fall of emergentism precisely to the success of quantum mechanics in explaining chemical forces. While I think it is true that the evidence cited by McLaughlin is sufficient to refute emergentism about chemical bonding, I do think it refutes emergence *simpliciter*. Whether or not there are emergent properties is not an all-or-nothing affair – the success of quantum mechanics tells us that certain properties are not emergent, but says nothing about the emergence, or otherwise, of properties that are, as yet, not quantum mechanically explicable. We

'multiple emergence' – if the emergent powers are possessed by different aggregates with nothing in common to them other than that they have this power, then we will need the emergent *property itself* to explain the emergent power. Suppose multiple emergence is true for mental properties. Individuals in the same mental state will have different physical properties, and will behave in a way that is not determined by the causal powers of those properties. In this case the mental property cannot be identified with any structural property. Rather, both physical and mental properties together determine behaviour. Let us call the first of our two versions of strong emergence the *strongly emergent powers* thesis; and the second the *strongly emergent properties* thesis.[221]

We turn now to weakly emergent properties, which have all the features of their *strongly* emergent cousins, except that their novelty does not violate $C_P$. This is a controversial move; if the emergent properties do not have effects that violate completeness, then how are their causal roles *novel*? The answer should be clear by now, for *general* novelty *is* consistent with $C_P$. By way of illustrating weak emergence, we will consider again the non-emergent property, temperature. The novelty of this property compared to particular aggregates of molecules with certain molecular velocities consists in its power to cause instances of other thermodynamic properties *whichever* structural property of the aggregate realizes it. As I have argued, such causal patterns as these are not duplicated at the microphysical level. But now take thermodynamics, and (i) make it functionally irreducible to statistical mechanics, (ii) weaken the strength of the modality in your preferred version of supervenience to nomological, (iii) let aggregates of statistical mechanical properties be sufficient for thermodynamic properties according to trans-ordinal laws that are not physically necessary. If the novelty of thermodynamic properties prior to the imagined

---

shall consider in chapter 7 whether there is any evidence against strongly emergent psychological properties.

[221] Strongly emergent properties will turn out to be important when, in 7.2, I appeal to strong emergence to discuss the putative evidence for $C_P$. My contentions there will be (i) that strongly emergent powers are inconsistent the completeness of physics, but not the completeness of the non-mental, (ii) that strongly emergent properties are inconsistent with both completeness theses, and (iii) that both strongly emergent powers and properties are open empirical possibilities given the available evidence.

transformation was (1) genuine and (2) consistent with $C_P$, then it remains so afterwards, despite that fact that by definition, thermodynamic properties are now weakly emergent.

A crucial fact about this conception of the novelty of weakly emergent properties is that the thing they do that is novel can only be specified if there are causal laws relating them to classes of effects, which laws can not be expressed without the emergent properties. It is the same as it is with (non-emergent) functional properties. Let functional property F's realization base be $f_1$, $f_2$, $f_3$,…,$f_n$ and functional property G's realization base be $g_1$, $g_2$, $g_3$,…,$g_n$. Suppose that on some occasion F is realized by $f_3$, which causes a realizer $g_1$ of G. If the novelty of F compared to $f_1$, $f_2$, $f_3$,…,$f_n$ is that an F has the power to cause a realizer of G *however F is realized*, then unless G *is* realized by $g_1$, there is nothing novel for F to do, for it will clearly not be a law that F-events cause $g_1$-events. Note that I do not intend talk of the general novelty of functional properties as in any way metaphoric: my contention, as I explained in 6.2, is that certain supervenient properties (e.g. functional properties) can capture generalisations that are missed at the physical (realizer) level. Now replace 'realization' with 'emergence' in the present example. What we have is two multiply emergent properties whose novelty consists in its being a law that F-events cause G-events *whatever base properties they emerge from*. Weakly emergent properties, then, possess general novelty but lack singular novelty. Notice that this sort of novelty depends on *multiple* emergence. It makes no sense to suppose that there could be a weakly emergent property with only one base property – for then the novelty of such a property's role compared to that of its emergence base could not be specified, and the base property could be substituted for the putative emergent *salva* the truth of any causal laws.[222]

Before proceeding to explain how weakly emergent properties cause a problem for the causal argument, I ought first to address a possible objection. The objection holds that

[222] On the assumption that substitution of co-referring terms is a permissible inference within the context of laws. I return to this issue in my discussion of the heterogeneity problem in 6.5.

weakly emergent properties as I have characterised them *are* functionally reducible, given the conception of functional reduction outlined in chapter 2. This presents me with a dilemma – either (i) weakly emergent properties are not emergent after all; or (ii) weakly emergent properties *are* genuinely emergent, but functional reduction as I conceive it is too weak to count as reduction. Here is how the objection proceeds. In 2.4, I argued that it is hugely implausible that the causal roles we use to functionally reconstrue a property M to be functionally reduced, can be specified in terms of properties of the reducing theory. Rather, the causal roles in question will be specified in terms of properties (M*, say) that *supervene* on properties in the reducing theory. Let P be M's emergence base, and P* be M*'s emergence base. Step (1) of the reduction process is fine, for we can treat M as a second-order functional property individuated by its causing M*. And M has a putative realizer in P. Now I said in 2.4 that functional reduction was a matter of finding implementing mechanisms for causal relations between supervenient properties, such as the one between M and M*; but why is this not possible if M and M* are weakly emergent? Why, in other words, is it not sufficient for the functional reduction of M that we explain how M's emergence base causes M*'s emergence base? The answer lies in recognising that M*, *qua* weakly emergent, supervenes on P* with nomological necessity. We cannot explain the occurrence of M* just by appeal to P* and physical laws; rather, we will need to appeal in addition to the synchronic bridge laws that govern M*'s emergence. It follows that we cannot explain how P plays the causal role individuative of M without appealing to emergence laws, which hold independently of the laws of physics. Now from this it follows that M is functionally reducible not to the physical, but to what we might reasonably term the 'nomological', or – perhaps better – the *natural*. If this is true, then indeed weakly emergent properties are nothing over and above the natural. But this is not inconsistent with their being emergent properties; nor is it inconsistent with the suitability of functional reduction for establishing physicalism. We just need to be careful about which laws and properties are included in the reducing theory.

If weak emergence is indeed a consistent possibility, then the causal argument for physicalism is not deductively valid. First, nomologically necessary sufficiency,

which is compatible with emergence, is clearly sufficient for non-coincidence. Second, while strongly emergent properties violate $C_P$, *weakly* emergent properties do not. And third, given that weakly emergent properties have general novelty in just the same way as supervenient but *non*-emergent properties do, weak emergence is every bit *as* consistent with $E_M$ as supervenience physicalism. The premises of the causal argument, therefore, are compatible with weak emergence, and so do not entail physicalism. A further argument against weak emergence is needed. As I said at the outset of this section, it is not my intention to defend weak emergence. In fact, I think that something is very badly wrong with it, and I will presently spend considerable time telling you what I think that something is. The fact remains, though, that whatever it is that is wrong with weak emergence, it cannot consist in being at odds with any of the premises of the causal argument. It's worth noting that it isn't all bad for the argument – the emergentist position with which it is consistent is quite close to physicalism; as we have seen, the two are often taken to have the same metaphysical commitments, differing only as to their epistemologies. Further, as we have seen, the argument establishes the falsity of substance dualism – for only if two particulars are "made of the same stuff" can one be synchronically sufficient for the other. Since we are already committed to the ontological independence of physical stuff in general, it follows that mental particulars and properties are ontologically dependent on physical particulars and properties. Emergentists agree, but maintain that contra-physicalism, some of the properties of physical things are something over and above the physical. Now this clearly means that there will be *emergent events*; mental events, although dependent on physical events, will consist in the instantiation of emergent properties such that the events would not occur but for the truth of trans-ordinal laws, and so fail to occur at worlds that are minimal physical duplicates of this one. I will now give two arguments against weak emergence, one epistemological, the other teleological.

### 6.4. An epistemic argument against weak emergence

This is a very simple argument that does not tell against the *notion* of weak emergence but rather against weakly emergent *mental properties*. It depends on an epistemological premise to establish its conclusion, which is that putative weakly

emergent mental properties would lack general novelty as well as singular novelty. But a property that lacks either kind of novelty is redundant; so weakly emergent mental properties are redundant. There is nothing, I should stress, in the *concept* of weakly emergent properties that precludes their having general novelty; however weakly emergent mental properties are not novel. They lack general novelty in virtue of the fact that their *explananda* (behaviours) are not weakly emergent. I do not think that this argument is particularly compelling; this does not concern me, however, as the teleological one to follow in 6.5 is much stronger. Here, briefly, is how the epistemological argument goes.

Suppose actual mental properties are weakly emergent. Now consider a minimal physical duplicate $w_d$ of the actual world $w_a$. By hypothesis, no weakly emergent mental properties will be instantiated at $w_d$. Unfortunately, our counterparts' bodies move in just the same way they do at $w_a$; they make noise, eat sausages, build walls, go to work, and so on. This is because the noises we make, and the walls we build, around here, are *not* weakly emergent – they really are nothing over and above the physical.[223] We know this because we have pretty good functional reductions of things like arm movements, houses, and the manner in which voice boxes succeed in making noises, to the physical. The problem for weakly emergent properties is that we attribute mental properties precisely in order to account for phenomena that are *common* to $w_a$ and $w_d$. But now it follows that any reason for attributing a mental property at $w_a$ will justify attributing it at $w_d$ too, as all the relevant *explananda* will occur just the same. We can appeal to mental properties at $w_d$ to predict and explain behaviour just as we can at $w_a$; it would be a miracle, then, if $w_d$ individuals did not actually *have* the properties attributed. But now it follows that *if* we are to maintain that there are weakly emergent mental properties at $w_a$, mental properties *here* must be instantiated twice over, one set of the non-emergent mental properties that are also instantiated at $w_d$, and one set of weakly emergent properties! Now that really is redundancy, for at $w_a$ we have both the weakly emergent ones that are missing from

[223] I note in passing that I am not convinced that this is true of sausages. Indeed, some of the sausages I have encountered have seemed to me to be beyond anything one might create by physical means alone.

$w_d$, and the functional ones that aren't, both doing exactly the same thing. Any laws that mental events enter into in virtue of their emergent mental properties, they will already enter into in virtue of possessing non-emergent mental properties. The problem is not, as I said, with the concept of weak emergence; rather, the problem is that for mental properties to be weakly emergent, we would need a set of events such that there are causal laws governing their occurrence that can only be framed in terms of weakly emergent properties. But there just isn't anything *missing* from $w_d$ that would serve to characterise the novel role of putative weakly emergent mental properties at $w_a$. The reason this does not count against the notion of weak emergence in itself is that it is not *a priori* that bodily movements, for instance, are not themselves weakly emergent. If they were, then there would be something for the weakly emergent mental properties to cause. But the actual world isn't like that, and the properties, the causing of which gives mental properties their characteristic causal roles, are properties that supervene with physical necessity.

To summarise, the *reductio* runs as follows. Suppose $w_a$ mental properties are weakly emergent. Suppose further, as seems scarcely deniable, that the interpretive practices we employ to attribute mental properties involve broadly causal-functional criteria. Any reason for thinking that mental properties are instantiated at $w_a$ is equally a reason for thinking they are instantiated at $w_d$ too. At both worlds, the attribution of mental properties has too much explanatory success to be false, so there are mental properties common to both worlds. Emergent mental properties, by hypothesis, are not instantiated at $w_d$. Therefore, if there are weakly emergent mental properties at $w_a$, mental properties *here* must be instantiated twice. Now appeals to epistemological principles in metaphysical arguments are controversial, and not without reason. What would be nice is if we could run an argument against weak emergence that did not rely on any such principles. Here is just such an argument, drawing on the so-called 'miraculous coincidence problem'. The argument is that in the case of weakly emergent properties, we lack an account of how it is that the same property gets to emerge from a variety of emergence bases.

## 6.5. A teleological argument against weak emergence

This argument, if cogent, will not show any inconsistency in weak emergence. Rather, it will show that supervenience physicalism has a distinct theoretical advantage over weak emergence. The advantage consists in the fact that certain *prima facie* problematic features of supervenience physicalism admit of a teleological explanation, whereas exactly analogous problematic features of weak emergence do not. In particular, I will argue that the so-called 'miraculous coincidence problem' can be solved teleologically for supervenience physicalism, but not for weak emergence. The miraculous coincidence problem is initially raised by Papineau as a problem for Fodor's antireductionist account of the relationship between sciences at different levels.[224] This problem is closely related to the debate between Fodor and Kim on special sciences, so a summary of the debate is in order before proceeding to the argument.[225]

Special science properties, Fodor claims, are both natural kinds and irreducible. Being a natural kind means you get to figure in laws; being irreducible means special science properties are not identical to physical properties, but supervene on them.[226] Anti-reductionists argue that the multiple realizability of special science kinds means no type identities, so no reduction. Reductionists counter that such kinds can be identified with, hence reduced to, the disjunction of all their possible realizers. Fodor argues against this kind of reduction on account of the heterogeneity of the properties in the realization base. The heterogeneity matters because, Fodor claims, the bridge laws required for reduction connect kinds to kinds, and a heterogeneous disjunction is not a kind. Kinds, for Fodor, are just the entities denoted by predicates that figure in laws, and Fodor claims that disjunctive predicates can't figure in laws. Kim replies –

[224] In Papineau [1985].
[225] Fodor's views on these matters can be found in Fodor [1974] and [1997], while Kim's most important contribution is (arguably) his [1992a].
[226] As before, this sense of 'irreducible' is not to be understood in terms of *functional* reduction. As I conceive of functional reduction, supervenient properties whose causal roles can be fully explained in terms of the causal roles of their subvening properties will count as reduced, despite the fact that they are not identical to their subvenient properties. In the sense of reduction involved in the Fodor-Kim debate, such properties are paradigmatically *ir*reducible. Nothing of import turns on such terminological matters.

quite rightly, in my view – that if disjunctions can't figure in laws, then neither can special science kinds, at least as long as these latter are conceived in such a way as to be necessarily coextensive with disjunctions. This dialectic leaves Fodor in a fairly unstable position – nobody really wants to deny that there are special science laws, but how *can* there be if special science kind terms are nomologically equivalent to disjunctions, and disjunctions can't figure in laws? The position in which it leaves Kim is little better. He too doesn't want to deny special science laws, but agrees with Fodor that disjunctions are not suitable for framing laws. As we saw in 2.3, Kim thinks that in virtue of the heterogeneity of the disjuncts, disjunctions of the realizers of special science kinds will not be projectible. To his credit, Kim sees that something has to give, and what gives for Kim is multiple realizability – he concludes that special science kinds figure in laws only to the extent that they *aren't* multiply realizable, hence not nomologically equivalent to disjunctions. The two horns of the dilemma, then, are Fodor's position, which seems to entail that there are no special science laws, and Kim's, which seems to embrace the type identity theory in its denial of multiple realization.

Fodor does not say exactly why *he* thinks disjunctions can't figure in laws; here is a brief recap of what he *does* say. Fodor claims that 'it's a law that…' is not a fully truth-functional context, meaning that (at least) some truth functional arguments are not permitted therein. Why does this matter? Well, one inference that won't be valid is presumably the inference from 'it's a law that $X \rightarrow Y$' & 'it's a law that $W \rightarrow Z$' to 'it's a law that $(X \vee W) \rightarrow (Y \vee Z)$'. If it were, then (since Fodor thinks natural kindhood is suitability for framing laws) we could *gerrymander new* kinds at will by creating new laws featuring them. So if we allow the validity of inferences such as the one above, then it seems, as Fodor says, that we have to give up the view that the kind predicates of a science are those that form the antecedents or consequents of its laws. It's not that any absurdities follow from allowing such inferences, but rather that they don't sit well with a prior theory about which predicates denote kinds.

The claim that the context 'it's a law that...' is intensional does rule out the creation of disjunctive kinds from truth-functions of existing laws. But it doesn't tell against the suitability of disjunctive predicates for framing laws *qua disjunctive*. This much ought to be clear from the fact that the invalidity of the stated inference form in the context 'it's a law that...' applies equally to gerrymandered *con*junctive kinds. Moreover, this line of argument does *not* depend on the level of heterogeneity of the disjunctions so formed. We certainly *could* gerrymander 'wildly' heterogeneous disjunctive kinds if contexts like 'it's a law that...' were *fully* extensional, but the manner in which they're *not* fully extensional tells equally against *all* gerrymandered kinds, regardless of how similar the original kinds are in their causal potencies. It seems, then, that Fodor's argument is aimed at *gerrymanderedness* rather than disjunctiveness or heterogeneity. But if this is so, then it will not apply to disjunctions that *aren't* gerrymandered, and disjunctions of the realizers of a given special science kind quite plainly aren't.[227]

For my part, I think that something like the following must have been implicit in Fodor's thinking on these matters. What if 'it's a law that…' is a fully *intensional* context? Consider the following inference form: from 'it's not a law that (F v G) → E' and 'necessarily H ↔ (F v G)' infer 'it's not a law that H → E'. If the context 'it's a law that…' is intensional, then quite clearly this inference is *not* valid. This would yield exactly the conclusion that Fodor wants, the view that the unsuitability of disjunctions for framing laws does not entail that special science kinds can't figure in laws. Fodor could maintain his antireductionism, via the thought that disjunctions aren't kinds, and escape Kim's objection that any problem for disjunctive laws is a problem for special science laws. However, the view that 'it's a law that…' is not a

[227] It should be noted that Fodor nearly admits as much, in his [1997] p.156, where he considers the possibility that some disjunctions may be projectible after all. Multiply realizable kinds, in contrast to gerrymandered ones, are coextensive with *open disjunctions*, in the sense that some of the disjuncts will be non-actual realizers of that kind. A closed disjunction, on the other hand, will be some finite disjunction such as 'Jadeite or Nephrite'. Talk of open and closed disjunctions clouds the issue somewhat, however, as this distinction does not track the all-important distinction between gerrymandered and non-gerrymandered. While I am happy to accept for the sake of argument that all closed disjunctions are gerrymandered, not all open disjunctions are non-gerrymandered. For instance, the disjunction of all possible realizers of all natural kinds is as open and as gerrymandered as it gets. A disjunction of all possible closed disjunctions is also open, as is the disjunction of any open disjunction with a closed disjunction.

*fully* truth-functional context must be sharply distinguished from the view that it's a fully *intensional* context. The problem in running the above line of argument is that Fodor has only given us reason to believe the former, whereas what we need is the latter. An example of a truth-functional argument that doesn't go through in the context 'it's a law that…' is sufficient to defeat the view that such contexts are *fully* truth-functional. However, *mutatis mutandis*, an example of a truth-functional argument that *is* valid within such contexts is sufficient to defeat the view that they are fully *non*-truth-functional. Can we give such an example? Apparently so. For instance, from 'it's a law that water at a pressure of 1atm boils at 100°C' and 'water = $H_2O$' we *can* infer 'it's a law that $H_2O$ at a pressure of 1atm boils at 100°C'. So substitution of co-referring kind terms looks OK. I can't think of an argument to block inferences such as this one, and clearly the burden of proof is on the proponent of the intensional view to provide one.

Let us grant Fodor that gerrymandered disjunctions are not suitable for framing laws. The crucial point to note now is that disjunctions of the realizers of functional kinds are *not* gerrymandered. Why? Because in order to *count* as realizers of a given functional property, all the disjuncts must play the causal role that defines it. This is where Papineau's [1985] argument comes in. If special science properties are multiply realizable (and so irreducible), then their realizers must be heterogeneous. But in that case, something has to *explain* how all the non-identical realizer properties at, say, the physical level, share the causal power constitutive of the functional properties at some special science level, say biology. Papineau turns Kim's argument on its head: Kim starts with the heterogeneity of the realization base and works 'bottom-up' to show that this heterogeneity leads to projectibility problems. Papineau starts with projectibility, and works 'top-down' to argue that heterogeneous properties play the same causal roles, which cries out for explanation. It would be miraculous if all the different realizer properties play the same causal roles by coincidence. Whence a dilemma: either there is an explanation of the otherwise miraculous coincidence, or special science properties are not multiply realizable after all. Papineau does believe in functional kinds, and offers teleological explanations of how it is they get to have

multiple realizations – the different realizers play the same causal role because they were *selected for in virtue of their causal powers*.[228] We already know that this is possible for artefacts – you can build a mousetrap out of just about anything, if you are clever enough, and have enough time on your hands; the explanation of how they all catch mice is teleological. That is what they are *meant* to do. Papineau's conclusion is that heterogeneity is no obstacle to lawlikeness provided the heterogeneous properties in question are selected in virtue of their causal powers. For Kim projectibility means uniform realization; for Papineau it means uniform realization *or selection*.

The problem all this generates for weak emergence is that there *could not be* a teleological explanation of how it is that the same weakly emergent property emerges from all the different physical properties in its emergence base. This is because although weakly emergent properties make novel patterns in relation to *each other*, without the sort of downward causal influence that strongly emergent properties have, there just isn't anything for any selection process to select *for*. Why would natural selection favour biological properties from which weakly emergent mental properties emerge over those from which no such properties emerge? The biological fitness of an organism depends on properties that are *not* weakly emergent. We can run a parallel argument to the epistemological argument of 6.4: in $w_d$, although the emergent properties are not instantiated, there are no corresponding differences in the fitness of any organisms that exist there. From this it follows directly, without appealing to epistemological premises, that if there are weakly emergent properties at $w_a$, then there is no teleological solution to the miraculous coincidence problem for these properties. It is important to note that the problem here is not that there is no functional reduction of the emergent properties to properties in their emergence base, for that will be common to *all* emergent properties. Rather, the problem is that there is no teleological explanation available of the otherwise miraculous fact that the same emergent properties emerge from a heterogeneous range of base properties.

[228] A similar line of response to Kim is to be found in Block [1997].

The difference is clear if we reflect on the fact that such teleological explanations are available for *strongly* emergent properties, despite the fact that their causal powers can not be explained in terms of those of the properties from which they emerge. A range of heterogeneous base properties could be selected for in virtue of being emergence bases for a strongly emergent property, as this latter will confer causal powers that are not conferred by its base properties. Weak emergence, on the other hand, is invisible to selection. Correspondingly, we are left with unexplained coincidences. What teleology offers, in essence, is the promise an explanation of how there could be irreducible patterns – in the case of weak emergence, it seems, we have the patterns without the explanation. Those who have no truck with teleological explanations, or think that they do not, in fact, solve the miraculous coincidence problem, are left with just the same sort of problem, and it is, I think, a far more serious problem that standardly acknowledged.[229]

If the preceding argument is cogent, then it seems Kim may be right in at least this much: if mental properties are emergent, then they had better be *strongly* emergent. But strong emergence is inconsistent with $C_P$ – so if $C_P$ is true, the combination of the causal argument with our two arguments against weak emergence entails physically necessary supervenience. Now, finally, we come to the question whether the evidence for $C_P$ is any good. Since both strongly emergent powers and properties are inconsistent with $C_P$, to the extent that the evidence for $C_P$ is good, it must also be good evidence against strong emergence. However, it is not, and so the evidence for $C_P$ is not good. My contention in what follows will be that the empirical evidence available to us at this stage tells against neither emergent powers nor properties.

[229] Jonathan Knowles is an example of one who recognises the problem but denies that the teleological solution works. See his [1999] for details.

# 7. Emergence and the Completeness of Physics

The purpose of this chapter is to show that current evidence does not support $C_P$. My argument will be directed at the non-triviality argument of 3.2 for the conclusion that completed physiology will not make ineliminable reference to mental properties. That argument, you will recall, is an induction from past successes in physiology – past successes in physiology have all involved entities like bones, muscles, neurones, impulses, neurotransmitters and tendons. But this, the argument of 3.2 went, gives us good inductive evidence that future successes in physiology will involve appeals to entities of a similar sort. And since *sui generis* mental properties are nothing at all like any of those entities, we may conclude that completing physiology won't require the introduction of mental properties. I will treat the conclusion of the non-triviality argument (that completed physiology will not involve *sui generis* mental properties) as of a piece with $C_P$. Nothing much turns on this for my purposes, as $C_P$ is the stronger thesis. My argument will depend on the possibility that mental properties are strongly emergent in the sense explained in 6.3. The burden of 7.1-7.3 is to argue that current evidence does not tell against the view mental properties are strongly emergent. Since strong emergence is inconsistent with $C_P$, if I am right that the evidence is consistent with strong emergence, then it follows that the evidence does not support $C_P$. If this is so, then there is a serious doubt as to the soundness of the causal argument. In the course of constructing my argument, I will explain what kind of evidence *would* support $C_P$. In particular, in 7.4 I argue that the evidence looks very much like a functional reduction of psychology to neurophysiology. But as we saw in 2.2, functional reductions establish supervenience without the need for argument. Correspondingly, I argued in 2.4, the reason we need a causal argument for physicalism just is that we don't yet have a reduction of mind. But this in turn means that evidence strong enough to support $C_P$ would establish physicalism about the mind directly, without the need for the causal argument. The soundness of the causal argument depends on good evidence for $C_P$; but good evidence for $C_P$ renders the argument unmotivated. First, then, let us take a look at the kind of evidence that we have at our disposal.

## 7.1. The nature of the putative evidence for $C_P$

In order to understand the problem, we will examine the story that Andrew Melnyk tells in favour of the completeness of physics.[230] Melnyk actually tells two stories, both of which we will consider. Melnyk argues for physicalism via a version of the causal argument, and endorses many of the theses I have thus far endorsed. His argument turns on versions of $E_M$, $C_P$ and a coincidence-based $O_D$. The difference between the way Melnyk runs the argument, and the way we have run it, is that rather than appealing to universal forms of $E_M$ and $C_P$, Melnyk appeals to particular existential instantiations of them. Nothing much depends on this difference, as it does not matter which particular events you pick. Like me, Melnyk observes that there are non-physicalist positions consistent with the premises of his argument, and like me, runs broadly epistemological redundancy arguments against these positions. Melnyk acknowledges that his causal argument is not deductively valid. In broad outline, his overall argumentative strategy is to compare the theoretical merits of physicalism and certain non-physicalist alternatives in explaining the otherwise coincidental occurrence of the mental and physical causes of bodily movements. I do not take issue with any of this; I am in fact in agreement with just about everything Melnyk says about what follows from the premises he endorses. As we saw in 6.5, weak emergence is at a distinct explanatory disadvantage compared to supervenience physicalism. While the problems we attended to there are not the same problems as those to which Melnyk draws attention, I think Melnyk and I can agree that given $C_P$, physicalism is the best way to make sense of the other premises. What I can't agree with, however, is that the evidence he describes supports $C_P$.

Melnyk asks us to consider a particular decision to clench your fist.[231] Fist-clenchings are constituted by contractions in the muscles of the forearm. The contractions of individual muscle cells we know, says Melnyk, to consist in "sliding, within each cell, of protein filaments of one kind over protein filaments of another kind." Now we also know empirically that the proximal causes of such slidings is "the *release of calcium*

[230] In his [2003].
[231] See his [2003] p.158ff for details.

*ions* from flattened vesicles that form a structure inside the cell called the sarcoplasmic reticulum." The point of all this is to establish that:

> (P1) Your decision to clench your fist caused…certain particular releases of calcium ions.[232]

Melnyk claims that no transmission principles such as those we rejected in 4.3 are required in order to conclude that this is so; rather, he appeals to the principle that correlations like the one between a decision to clench your fist and the release of calcium ions are best explained by positing a causal relationship between the correlated events. Proposition (P1) is the $E_M$ of Melnyk's argument. Clearly his intention here is to set the stage for causal competition between the decision and the physical causes of the Calcium ions, for which he argues next. Now, as we saw in 4.4, mental and physical causes need not compete for the very same effects in order for their co-occurrence to be a coincidence that stands in need of explanation. No matter; let us grant for the sake of argument that (P1) is true. Melnyk claims that there is empirical evidence that weighs in favour of the proposition that

> (P2) There were sufficient *physical* causes for the particular releases of calcium ions mentioned in P1.[233]

This is the crucial premise, and plays the same role in Melnyk's particularised causal argument as $C_P$ does in the more familiar version. It is no coincidence that of the two stories that Melnyk tells about why we should believe P2, one is very similar to the non-triviality argument for $C_P$, and the other is an explicit argument for $C_P$ from which, clearly, P2 follows *a fortiori*.

The first story Melnyk tells about why we ought to believe P2 is, in my view, very close to the non-triviality argument of 3.2 for the completeness of physics. I should point out, however, that Melnyk at no point endorses that argument, and takes his first

[232] Melnyk [2003] p.158.
[233] Melnyk [2003] p.160.

story to be an argument only for the particular claim expressed in P2. Our best theories, he claims, show that we can trace the causal ancestry of the release of calcium ions back into the brain. The following statement is illuminating:

> The releases of calcium ions…are phenomena whose biochemical causal antecedents can be traced…first to activities in the motor neurones that innervate the muscle, and then to activities in other neurones that interact with motor neurones, and so on back into the brain as far as you care to go; the reason for thinking this tracing to be possible is that neuro-anatomists have actually traced the pathways of bundles of neurones into and out of the brain, *and the biochemistry of the individual neurones that make up these bundles is well understood*.[234] [My italics.]

The 'well-understood' in the italicised passage means well understood *physically*. We have good explanatory physical accounts of how individual neurones work in physical terms – although Melnyk does not mention it, there is a branch of physics know as 'Biophysics' that specialises precisely in accounting for things like neuronal firings in terms of physical quantities like charge and processes such as ionic diffusion. The argument now is this: (i) we can trace the causal ancestry of the calcium ions that proximally cause muscle movements back to bundles of neurones; (ii) the functional properties of individual neurones is well-understood in physical terms; so (iii) the release of ions has a sufficient physical cause. Now Melnyk admits that we do not have anything like a complete understanding of each process involved in causing the ions to be released. The central feature of the story, I suggest, is that nothing in it appeals to *sui generis* mental properties, or anything like them. Trace the ancestry of the fist-clenching back as far as you like; nothing in what we *do* know about the chain of causes seems to require that any of them be irreducibly *mental*. Understood in this way, Melnyk's line of argument is clearly very close to the one we considered in 3.2 in support of $C_P$. For in essence, Melnyk's claim is that however far back we look into the causal ancestry of the particular release of Calcium ions under scrutiny, we find only biochemical causes. Melnyk goes further, arguing that the biochemical causes are themselves physical via the premise that biochemistry is functionally reducible to

[234] Melnyk [2003] p.160.

physics, but as we saw in 3.2, this is unnecessary. If our current best understanding of the causation of bodily movements involves only entities of a similar kind, none of which are anything like *sui generis* mental properties, then it appears that the non-triviality argument can induce that completed physiology will be non-mental in character, which is all we need in order to render the completeness of *physics* non-trivial. If muscle movements have causes that don't look *sui generis* mental, then *a fortiori* we can be relatively certain that no *physical* effects are the results of *sui generis* mental causes either.

The second line of argument that Melnyk offers in favour of P2 focuses directly on the successes of physics in order to establish $C_P$. This is very similar to the line of argument we rejected in 3.2 as containing a sampling error. This one contains such an error as well, as we shall see. Melnyk has this to say about the successes of physics:

> [C]urrent physics' success to date in finding that *many* physical events have sufficient physical causes provides inductive evidence that all physical events, including both unexamined physical events and examined-but-as-yet unexplained physical events, have sufficient physical causes.

It is difficult to see how any of this could be relevant to the matter at hand. Nothing in what a Cartesian dualist has to say involves the movements of atoms in cloud chambers being caused by special mental forces. Complete physical explanations of what goes on in particle accelerators are perfectly consistent with the *in*completeness of physics, due (for instance) to special mental forces that cause *behaviours*. It looks as thought the direct argument from physics will fail to convince anyone who isn't already convinced by $C_P$, for the doubters think that there's something special about what goes on in brains that involves non-physical forces. What goes on in particle accelerators, while interesting, is by the by. Melnyk considers this problem, and offers the following rebuttal:

> Current physics shows no sign at all that contemporary physicists expect to find any physically anomalous phenomena whatever inside human brains, which seem, from the physical point of view, to be quite unexceptional…[the biochemistry of brain cells] is apparently no different from that of cells of

> other types; likewise, presumably, for their physics, given the physical realization of biochemistry that I am assuming….[235]

The argument, if I understand it correctly, is this: we understand the biochemistry of individual cells to the point where we can be confident their operation is wholly explicable in terms of physical laws and properties. Let us grant Melnyk for the sake of argument that individual nerve cells are functionally reducible to physical properties and laws. This is actually quite plausible. The property of being a neurone is plausibly functionalizable, multiply realizable (in the same way, perhaps, that we found temperature to be multiply realizable in 2.3), and physically realized. Biophysics offers pretty complete explanations of how the physical realizer properties of neurones get to play the causal roles that (given functionalization) individuate the property of being a neurone. But now, the argument goes, brains are just big bunches of these cells stuck together, so we should expect what brains do to be explicable in physical terms as well. So much for the evidence; we turn now to the question whether it is any good.

### 7.2. Evaluating the evidence

We will take Melnyk's second argument first. Does the fact that the *parts* of brains operate according to physical laws entail that *whole* brains operate according to the same laws? The problem that the appeal to the 'physically unexceptional' nature of brain cells is supposed to solve is that the success of physics in explaining why atoms behave as they does not entail that it will have similar success in explaining why brains behave as *they* do. We can not induce from the fact that isolated atomic events have sufficient physical causes to the conclusion that *behaviours* have sufficient physical causes, for the sample in question just isn't of the right sort. But now it is unclear how the appeal to brain cells is supposed to help. Why not just hold that since brains are fully composed of atoms, which are physically unexceptional, it follows that we should expect no physical anomaly in the behaviour of brains? The problem here is that what strong emergentists endorse is *precisely* the claim that certain

[235] Melnyk [2003] p.161.

configurations of atoms give rise to properties whose powers are not determined by those of their physical base properties. In other words, even though atoms behave in one way, if you put enough of them together, properties emerge that possess *singular novelty*. This, you will recall, is defined in terms of there being a token event whose occurrence can not be explained without reference to the emergent property. We will have more to say about this matter in 7.3; for now, notice that the appeal to brain cells adds nothing. If mental properties are strongly emergent, then no *single* brain cell is going to be nomologically sufficient for them; rather, *aggregates* of brain cells possessing certain structural properties will be the minimum units of emergence. So again, the appeal to the functional explicability of brain cells in physical terms will not convince anyone who isn't already convinced, that brains do not possess emergent mental properties. The sampling error in this case consists in inducing the functional reducibility of brains to physics from the functional reduction of neurones to physics. Nobody ever held that the effects of individual neurones taken in isolation were the sort of things we needed *sui generis* mental forces to explain. The argument direct to $C_P$ from the success of physics fails, for just the same reasons as those given in 3.2; that brains are 'physically unexceptional' is precisely what an emergentist will deny.

Let us return, then, to the much more promising evidence from physiology. In what follows, I will argue that everything Melnyk says about the causal ancestry of bodily movements is something a strong emergentist about mental properties should agree with. Recall from 6.1-6.3 that I think of strongly emergent properties in the following way: when you put enough of a certain kind of part together in the right way, the composite whole you get behaves in a way that isn't determined by the causal powers of the parts. Rather, the causal powers of such aggregates are determined by the parts along with the emergent property itself. In the present case, we can think of it like this: strong emergence is the hypothesis that explaining the causal powers of brains involves ineliminable reference to mental properties. This seems a perfectly legitimate hypothesis – for it is clearly not *a priori* that all composite entities behave in a way that *is* determined by the laws that govern their parts. Allow for the sake of argument that Melnyk is right that biochemistry reduces to physics; from this it follows that the

parts of brains obey physical laws. What strong emergentists will deny is that the causal powers of *whole* brains are determined by their physical properties and laws of physics: there are physically possible worlds in which brains behave in a quite different way to the way they behave around here.[236] Now as we saw in 6.3, this licenses two possibilities, which I termed the strongly emergent powers thesis, and the strongly emergent properties thesis. To recap, if all emergence bases of an emergent causal power have the same structural property, then the novel powers are properly conceived as powers conferred by the structural property itself (in other words, *the structural property has emergent powers*). If, on the other hand, there are multiple emergence bases for an emergent power, then we will need to invoke an *emergent property* to explain it. Now as I said, each of these theses seems to me to represent a perfectly plausible empirical possibility. Why *should* composite events cause just what you would expect them to given their constituents? And why, if we accept multiple realization as not only possible but actually quite likely, should *different* aggregates of events not exhibit the same non-physically determined behaviour, in virtue of a common emergent property?

Before proceeding, a quick note is in order on the relationship between the two emergence theses and the two related completeness theses discussed in 3.2, viz. $C_P$ and the completeness of the non-mental. If the emergent powers thesis is true, then there will be no complete physical explanations of bodily movements without appeal to the structural properties of the aggregates from which mental powers emerge. But these structural properties look nothing at all like anything in current physics. It is possible to argue for their inclusion in completed physics, but to my mind this breaks the inductive link between current and future theory upon which the non-triviality argument rests. Still, the structural properties in question will be non-mental; all that the emergent powers thesis claims, on this account, is that there are some non-mental structural properties whose powers outstrip the powers of their constituents. Physical

[236] Kim runs the supervenience argument against emergence to show that even if strong emergence is true, still the causal powers of things with strongly emergent properties are determined by their emergence bases, and that as a result, the efficacy of the emergence base properties pre-empts that of the emergents. We address this argument in 7.3 below.

science broadly construed, whilst arguably rendered incompletable with just *physical* entities, will nonetheless be completable by the incorporation of the structural properties that mental powers emerge from. This, as I said in 6.3, is (arguably) the kind of thing Broad had in mind when he claimed that certain chemical properties were emergent; but all that follows from this view is that chemistry and physics are arranged *horizontally* and are mutually supporting, rather than arranged vertically in a hierarchy, as reduced and reducing theories respectively. But now look what happens if the strongly emergent *properties* thesis is true, and the only entities available to explain the emergent powers are the mental properties. If this is the case, then physics will not be completable without the incorporation of *sui generis* mental properties. Now if that is how things are, then clearly supervenience of the mental on the PHYSICAL will be trivial, which is why we need the non-triviality argument to rule out positions like the emergent properties thesis. In the remainder of this section, and the next section, we will see exactly why the argument fails to do so.

Consider again Melnyk's causal ancestry argument. Strong emergentists will agree, in the first instance, that fist-clenchings have sufficient physical *proximal* causes, for no one could reasonably suppose that mental properties or powers emerge from the properties of your *hand*. The emergence bases for strongly emergent properties are most plausibly brain properties, and this being the case, emergentists will continue to agree with Melnyk until we trace the ancestry of the clenching back to the brain. This is a crucial, but easily overlooked, point. Imagine, for the sake of argument, a Cartesian spirit operating a robot by remote control; undeniably, the ghost causes the robot's movements. In the imagined case, the robot's motion has both physical and non-physical causes, occurring at different stages in the same causal chain. Emergentists will not deny that emergent properties cause physical events that do not possess emergent properties, and whose subsequent effects can be fully explained by appeal to physical laws. The brain is supposed to be where the emergent properties emerge, and the point at which physical events occur that do *not* have complete sufficient physical causes. But even there, the disagreement is not as stark as it may initially appear. For on the conception outlined above, strong emergentists do not

need to deny that the effects of emergent events *have* physical causes; what they will deny is that those effects have *complete*, *sufficient* physical causes. Nothing in the way I have set up the strongly emergent properties prevents the view that it is the physical emergence base events *together* with the emergent event that are causally sufficient for the relevant effects.[237] But if this much is granted, then a strong emergentist should agree with everything Melnyk says!

Consider: since emergent properties supervene on physical properties, emergentists must hold that you can trace the causal ancestry of a mentally caused bodily movement back to the physical emergence base of the mental event that causes it. The locus of the disagreement between $C_P$ and emergence is that the former affirms, whereas the latter denies, that the physical base events are complete, sufficient causes of the movement. It as unfortunate tendency of physicalists to speak as though the *in*completeness of current theory does not impact on the question whether current (limited) explanatory success entails future completability. Melnyk, for instance, agrees that we do not have anything like a complete understanding of how neurones interact with each other to cause behaviour: "[O]ur biochemical understanding of the causal ancestry of calcium ion releases is certainly not complete."[238] The tendency is unfortunate because what an emergentist will deny is not that *in*complete causal explanations of bodily movements can be given in non-mental terms, but that such movements have complete, sufficient physical causal explanations!

Nothing in the strongly emergent properties thesis precludes the view that the causes of behaviour can be traced back to physical antecedents – in fact, emergence requires that they can be so traced. What emergence denies is that behaviours have complete,

[237] Lowe [1993] thinks of the relationship between the causal powers of emergent properties and those of their emergence bases in a similar way. Emergent properties are best thought of, he claims, not as initiating the causal chains that culminate in bodily movements, but as "inducing certain patterns of convergence amongst neural events," (p.638). Emergent properties exert a co-ordinating downwards influence on their otherwise unrelated base events; without such an influence, Lowe suggests, the fact that such unrelated events manage to combine to cause behaviours would be a coincidence. Lowe admits that this speculation is open to empirical disconfirmation – but like me, does not see anything in current science to exclude its possibility. What is important for my present purposes is that on Lowe's model, it is mental and physical causes *in combination* that are the sufficient causes of behaviour.
[238] Melnyk [2003] p.161.

sufficient physical causes. But if this is so, then how could it possibly count *against* emergence that we are capable of giving physiological or biochemical explanations of behaviour which, by admission of even the most diehard physicalists, are far from being complete? Consider again the non-triviality argument, and focus on the sense in which past physiological explanations have been *successful*. As the evidence cited by Melnyk clearly shows, we have not yet been successful in giving complete explanations of bodily movements in physiological terms. Of course, at certain stages complete explanations can be given, for instance the explanation of the causal relationship between Calcium ions and muscle contractions. But that is a (relatively) complete explanation of the (relatively) *proximal* causes of those movements. As our tracing of neural pathways takes us back into the brain, the explanations become more and more incomplete. Now we can all agree that our limited success in this area has been achieved without *sui generis* mental properties. However, no induction from past *partial* explanatory success without mental properties to future *complete* success is going to work – for the absence of mental properties from the account might be just what is missing, and the reason why the explanations aren't complete! Clearly, then, the non-triviality argument too contains a sampling error. What the sample ought to contain is a stock of (at least *fairly*) complete non-mental explanations of effects we know to have mental causes. What it actually contains is nothing of the sort. Our stock of successes in providing incomplete physiological explanations of the causation of bodily movement is suitable only for inducing that we will, in future, be able to provide further *partial* explanations of such movements in the same, or similar, terms. The contrast with physics is a stark one indeed; the reason we are right be confident that no scientist will ever have to leave the physical realm to explain what goes on in cloud chambers and particle accelerators, is that we have an excellent stock of relatively complete explanations of such phenomena in physical terms.

In the next section, we will appeal to an argument of Kim's against emergence to show how that, in fact, strong emergence is also consistent with our being able to give very good causal explanations (much better than the ones we have at present) of the effects of emergent mental properties without mentioning those properties at all. The

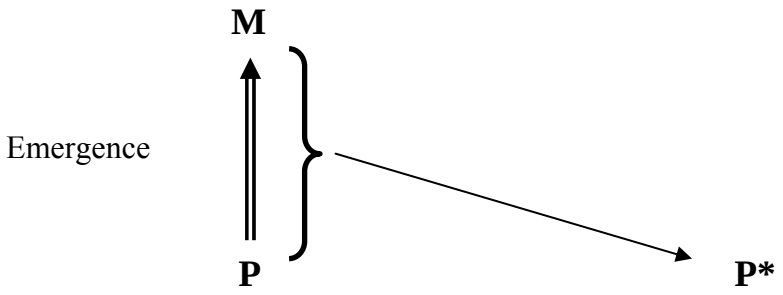
mere fact that a property is not mentioned in such an explanation, as we shall see, does not entail that the property lacks singular novelty with respect to those effects. Through this, we will give an account of what the evidence would have to look like in order to support the view that there are no strongly emergent properties, and so provide crucial empirical support for $C_P$.

### 7.3. The supervenience argument again

Recall Kim's argument against supervenient causation, which we explored in 5.1, and refer back to the upwards-downwards transmission diagram. Kim runs exactly the same two-stage argument against emergence.[239] Emergentists need to endorse mental-to-mental causation from M to M*, which presupposes downwards causation from M to P*. But then why doesn't P, as M's emergence base, pre-empt the causal status of M? As we saw, Kim's argument involves 'pushing causation around' – downwards from M to P*, and also upwards and then downwards again from P to M to P*. Since I already accept that strongly emergent properties have singular novelty, and that this amounts to diachronic downwards causation, I will spare Kim the downwards transmission argument, and focus exclusively on the upwards-downwards transmission argument.[240] As before, this argument purports to show that M's causal efficacy is pre-empted by that of its base property. Allow that strongly emergent properties combine with their emergence bases to cause physical effects. Forget about mental to mental causation. We can draw the situation like this:

[239] He uses the downward causation argument against emergence in his [1992b] and [1999a]. The argument has exactly the same form whether Kim is arguing against emergence or supervenience physicalism. The sole difference is that when arguing against emergence, Kim rightly does not appeal to $C_P$. In what follows, I will show that this difference is crucial.

[240] Strictly speaking, I should not call it the upwards-downwards transmission argument in the present context, for there is no downwards transmission part. But it is only because, as I made clear in 6.2, I grant Kim that singular novelty entails downwards causation, that the downwards transmission argument is omitted in the present case.




Now Kim's argument against supervenience, you will recall, was that it renders mental causation unintelligible. He runs a similar argument against emergence, only this time the charge is inconsistency. Strongly emergent properties have novel causal powers, says Kim. But now allow that causation is to be understood as nomic sufficiency. P is nomically sufficient for M, and the joint occurrence of P and M is nomically sufficient for P*. But now it follows that P, as M's emergence base, is nomically sufficient for P*, and so pre-empts the alleged singular novelty of M. So contrary to supposition, emergent properties do not have novel causal powers. Here is how Kim puts it:

> The critical question that motivates the argument is this: If an emergent, M, emerges from basal condition P, why can't P displace M as a cause of any putative effect of M? Why can't P do all the work in explaining why any alleged effect of M occurred?....Now we are faced with P's threat to pre-empt M's status as a cause of P*….For if causation is understood as nomological (law-based) sufficiency, P, as M's emergence base, is nomologically sufficient for it, and M, as P*'s cause, is nomologically sufficient for P*. Hence, P is nomologically sufficient for P* and hence qualifies as its cause….This appears to make the emergent property M otiose and dispensable as a cause of P*; it seems that we can explain the occurrence of P* simply in terms of P, without invoking M at all. If M is to be retained as a cause of P*…a positive argument has to be provided, and we have yet to see one. In my opinion, this simple argument has not so far been overcome by an effective counter-argument.[241]

This argument, if cogent, will show that the putatively novel causal powers of strongly emergent properties are in fact conferred by their emergence base properties. In effect, the argument is that since P contributes M, then any causal powers contributed by M will be contributed by P as M's emergence base. I do not find this argument at all compelling, and will give two responses. My first response is that

[241] Kim [1999a] p.32.

when the upwards-downwards argument is directed against *emergence*, it lacks a crucial justification, in the form of $C_P$, for the conclusion that P pre-empts M's efficacy. Kim fails to fully appreciate just how much of a gap in the argument $C_P$ leaves. My second response in to argue that there are good general grounds for resisting the view that P has the causal powers of M.

*First response*. I do not deny that P is nomically sufficient for P*. Further, I am prepared to grant Kim for the sake of argument that this makes P a cause of P*. What I deny, however, is that either of P or M *on its own* qualifies as a complete, sufficient cause of P*. As I stressed in 6.2, the correct interpretation of the emergentist picture is that M and P *together* suffice to causally determine P*. Now crucially, when Kim runs this argument against supervenience physicalism, he argues that we cannot view P and M as jointly causing P*, for in that case we should have a physical event (P*) with only a partial physical cause. The problem with this view, as we saw in 5.1, is that it is inconsistent with $C_P$ that the causation of a physical event should require a *non*-physical event M in addition to P. But then what makes M dispensable as a cause of P* is precisely that P* has a complete, sufficient physical cause, and this is exactly what strong emergentism denies! It follows that Kim's central reason for not taking M and P to jointly determine P* is missing when the subject of the argument is strong emergence. We can put the same point slightly differently, for the sake of clarity. The upwards-downwards argument, directed against emergence, is supposed to *show* that P has any putatively novel powers contributed by M, hence that upwards determination is inconsistent with downwards causation. But appealing to $C_P$ in this connection would merely beg the question – if M exerts a downwards influence without which P* would not occur, then $C_P$ is false. The argument, if successful, ought to *entail* $C_P$, despite the putative singular novelty of M.

Kim of course realizes all this, which is why he does *not* appeal to $C_P$ in the quoted passage. But my point now is that shorn of $C_P$, the argument is really rather weak. Three short arguments can be run to undermine it. Since an appeal to $C_P$ is out of the question, Kim must maintain that nomic sufficiency is sufficient for *sufficient*

causation, in order to conclude that P is causally sufficient or P*. Now we have the usual overabundance of causes for P*. The first short argument is this: unless we can appeal to $C_P$, why choose P over M as P*'s cause? Why not the other way around? And the second short argument: another thing strong emergentism denies (by implication) is that nomic sufficiency is sufficient for sufficient causation. P, as M's emergence base, will be nomically sufficient for P*, and also (I grant for the sake of argument) a cause of P*; but P will not be a *sufficient* cause of P*, precisely because of the singular novelty of M with respect to P*. P is best seen as a *part* of the sufficient cause of P*, the other part of which is M. And the third: nomic sufficiency alone is not sufficient even for *causation*, let alone complete, sufficient causation. It is well-known, for instance, that effects of a common cause are nomically, but of course not causally, sufficient for each other. What is missing is an argument from Kim to the effect that the nomic sufficiency of P for M is sufficient to allow P to displace M as a cause of P*. At the end of the quoted passage above, Kim says: "In my opinion, this simple argument has not so far been overcome by an effective counter-argument." I have a counter-argument to Kim's simple argument. It is simplicity itself, and it goes like this: *what* argument? I note in passing that there is a sense in which we can explain the occurrence of P* without reference to M. P's nomic sufficiency for P* means that we can frame explanatory laws covering P and P* that do not mention M at all. As Lowe [2003] argues, the nature of emergence makes possible certain contexts of explanation in which the emergent properties (although in possession of singular novelty) are *invisible* to investigators. This point is of crucial importance, and we return to it presently.

*Second response*. This is not a direct response to any argument of Kim's, but is rather an attempt to give content to the notion, crucial to my conception of emergent singular novelty, that emergent properties contribute causal powers over and above those of their emergence bases. Now P above will be a complex physical event – or if you prefer, an aggregate's exemplifying a structural property at some time. While I agree with Kim that P is nomically sufficient for P*, I disagree that this violates the singular novelty of M. We can give *metaphysical* content to M's novelty via the

thought that the causal powers of M are not *physically* determined by the powers of P. That is, if there are physically possible worlds at which M is not instantiated and aggregates such as P do *not* cause P*, then M's novelty will be intact despite the fact that P is nomically sufficient for it. If our best physical theories about the causal powers of events of the same type as P, tell us that in fact we should expect them *not* to cause events of type P*, then that gives us a *reason* to think that M is contributing a power not physically determined by P. That is, if the power of P to cause P* persistently resists functional reduction of P to its components, then eventually we may be forced to conclude the singular novelty of M in this case. We would be inclined to make such a decision, I suggest, if our theories about the components of P could explain their behaviour in isolation very well, but consistently fail just when such components are put together in such a way that M is instantiated. In less abstract terms, we can put the point like this: if biophysics fails to explain the behaviour of neurones *just when those neurones are arranged in such a way that they instantiate a mental property*, then it will look as though the mental property is strongly emergent. So much for Kim's argument against the singular novelty of strongly emergent properties. What I want to take from it, however, is the point that the nature of emergence entails that despite the fact that M is necessary for P*'s occurrence, as M's emergence base, P is nomically sufficient for P*. This has extremely important consequences for the nature of the evidence that would be sufficient to support the truth of $C_P$.

Refer back to 2.5, and grant for the sake of argument that we are in fact a bit further advanced scientifically than I suggested we are. Specifically, grant the following steps in the functional reduction of mind are complete: (i) we have completely functionalized the mental properties, and so know exactly what causal powers their putative realizers will need to have; (ii) we have located all the (putative) possible physiological realizers of mental properties, and as we thought, multiple realization is true. Now let the realization base of M be $P_1$, $P_2$,…,$P_n$. Let $vP_i$ represent the disjunction of all the Ps. Either the M is strongly emergent with respect to $vP_i$, or it is not. Given that nomic sufficiency holds between each $P_i$ and P* even if M is

emergent, it follows that $\{vP_i\} \rightarrow P^*$ is a law regardless of whether or not M is emergent. It follows that even if mental properties are strongly emergent, we will be able to express laws relating their emergence bases to their effects, *laws that do not mention the emergent properties at all*. But on the reasonable assumption that at least some kinds of explanation involve subsumption under a law, it follows that we can explain the occurrence of P* without mentioning M, despite M's singular novelty in causing it. Further, these laws will enable us to predict bodily movements on the basis of physical properties alone. Why am I telling you this? Well, the successes Melnyk cites in support of $C_P$ are nowhere near as impressive as the kind of successes we would have in the hypothesised situation. Other than irrelevant successes, all Melnyk *really* cites in favour of $C_P$ is the traceability of causal chains that culminate in bodily movement back into the brain. I do not dispute that this is an interesting fact; but if explanation by subsumption and non-mental prediction of motion are compatible with strong emergence, then what hope does *this* fact have of supporting $C_P$? Of course, stage (iii) in the functional reduction is the crucial one ontologically, for as we have seen, this will consist in providing a physiological *explanation* of how each $P_i$ gets to cause P*. And this, presumably, will involve explaining how things with the physiological properties of the Ps play the causal role associated with M, which involves, *inter alia*, the power to cause P*. In other words, showing that a complex state P plays the causal role of M will involve showing, given the laws that govern the components of P, that P causes P*. And this is precisely what emergence claims we won't be able to do, for part of M's being strongly emergent is that we would expect things with the structural properties of the $P_i$ *not* to cause P*.

What the above argument shows is that nothing in the theory that mental properties are strongly emergent precludes our being able to frame causal laws relating neurophysiological event types to bodily movements, without mentioning the emergent properties at all. But as I said, on the view that explanation involves nomic subsumption, it follows that even our ability to give relatively *complete* causal explanations of behaviour in terms of complex structural physiological properties would not discriminate between physicalism and emergence as metaphysics of mind.

The differences between these two metaphysics only begin to show when you try *reducing* the explanatory laws to laws governing the behaviour of the parts of the Ps above. Emergentism disagrees with physicalism in that it denies that the causal relations certain complex events enter into can be explained in terms of laws governing their parts. In the case of strong emergence, this means that the differences show up when we try to reduce $\{vP_i\} \rightarrow P^*$ to more basic laws, for instance the laws Melnyk discusses according to which we explain the behaviours of *single* brain cells. The reason why there is no evidence for $C_P$, in a nutshell, is this: we have yet to even discover the $P_i$ such that it is a law that $\{vP_i\} \rightarrow P^*$, and yet evidence against emergence only begins to accrue as we manage to explain *how* the Ps play the causal roles they do in terms of the laws that govern the behaviour of their components.

It will help to cast the argument in less abstract terms. Let M = the desire to sip my wine, and its putative realizer P be an aggregate of individual neuronal events; let P* be my hand grasping my glass. We can predict and explain the grasp from the occurrence of P, without mentioning M, but that does not settle any matters of metaphysics. The central metaphysical question here can be stated thus: is P a realizer, or an emergence base, of M? The matter can only be settled in favour of the realization metaphysic by *showing* that this instance of P realizes this instance M. But that involves showing that things with P's structure have the power to cause this instance of P*. The deduction clearly achieves nothing if it relies on M's power to cause P*, for this too is consistent with strong emergence. Rather, we must deduce M from laws that govern the behaviour of P's component neurones. And that, of course, is just a functional reduction of this instance of M to this instance of P.

### 7.4. Prospects for the causal argument

The upshot of 7.1-7.3 can be stated as follows. The evidence for physicalism cites incomplete explanations of behaviour in physiological terms as evidence for the completability of physiology without *sui generis* mental properties. However, emergence does not preclude the possibility of such incomplete explanations as these – emergence instead claims only that mental properties have singular novelty not

determined according to physical law, so that no *complete* explanation of behaviour is possible in non-mental terms. The non-triviality argument contains a sampling error – the successes just aren't of the right kind, as any good emergentist will agree that incomplete explanations of behaviour can be given without citing mental properties. In fact, as we saw in 7.3, strong emergence even allows for *causal laws* relating the properties in the emergence base to the characteristic effects of the emergent properties, laws that do not mention the emergent properties at all. All that emergence claims is that the functional reduction that begins with finding the law $\{vP_i\} \rightarrow P^*$ can not be finished. Since we haven't even found the $vP_i$ yet – the determinants (be they emergence bases or realizers) of mental properties – it is wholly unrealistic to suppose we can prejudge on the strength of what little we *do* know that *whatever* these determinants turn out to be, their causal powers will be reducible to the powers of their non-mental parts. Whence a variation on Hempel's dilemma for $C_P$: nothing in the evidence discriminates between (i) that *sui generis* mental properties *will* be needed in order to explain the casual powers of the $P_i$, and (ii) that such properties will *not* be so required. The non-triviality argument fails.

Let me be clear about this much: I do *not* think that it is impossible to find out empirically that physicalism is true. I just think it will take considerably more work than most physicalists are prepared to accept. Here is the work that it will take. Having discovered the minimum physical base properties (the minimum units of determination) for mental properties, you then have to proceed to explain how they play the causal roles you used to characterise the mental properties, without, of course, appealing to the mental properties to do so. One possibility, of course, is that when we discover the physical base properties, we find that everyone with a given mental property shares a structural property specifiable in wholly physiological terms. Now if this is so, it matters not whether the structural property has emergent powers – for as we saw in 7.2, even if it does, we can appeal to the property itself to explain the powers. Physicalism (or at least, non-mentalism) can quite happily accommodate the view that when you combine physical entities so that they have a specific structural property S definable in physical terms, they behave in a way that is neither

determined by, nor (it follows) predictable by reference to, the S's constituent physical properties. Whether or not we all do share a physically definable structural property when we share a mental property is, it goes without saying, an empirical matter; one which, again, current evidence does not decide.

On the other hand, if things turn out as I described in 7.3, then in order to gather evidence that $C_P$ is true, we must continue with our reduction. In particular, we will need to show that the Ps that determine M do so just in virtue of being the P-events they are, which in turn will be a matter of appealing to more basic laws that govern the behaviour of the events that compose the $P_i$ in order to show that, pace emergentism, causing P* is just the sort of thing we would expect them to do given their neurophysiological constituents. This is the only way to support the conclusion that mental properties are not emergent – for it's hardly as though you can subtract the mental properties from the $P_i$ to see if the they behave in the same way on their own! (By hypothesis, each $P_i$ is sufficient for M; the question concerns the strength of sufficiency. Is it physical, or nomological?) What we can do, however, is try to explain the behaviour of M's putative realizers in terms of simpler properties that *don't* realize mental properties. As we saw in chapter 2, if we can explain the causal powers of the $P_i$ in the non-mental terms of some reducing theory, then M will count as functionally reduced to that theory and its properties, relations, and so on. Reduction of some form, I maintain – be it functional or not – is the only method at our disposal to show that M lacks singular novelty. If we can explain the effects of an M-instance wholly in terms of a $P_i$,-instance, then by definition the best kind of novelty M can hope for is general novelty. So evidence for $C_P$ is possible, it just needs a physiological explanation of how the physiological base properties of M play the causal roles that are required to count among M's realizers.

Unfortunately, if all this is correct, then the prospects for the causal argument are bleak. For if the Ms are functionally reduced, then we have no need of the argument in the first place! The argument goes in a circle: (1) we need the argument because we lack a functional reduction of mind; (2) the argument can't be run without $C_P$; (3)

evidence supports $C_P$ just in case it counts against strongly emergent properties; (4) the only evidence that will count against strong emergence is a functional reduction of mind. I anticipate an objection to (4), and I will take a moment to explain and rebut it. Surely, the objector asks, you demand too much? In support of Melnyk, one might argue as follows. Melnyk does not claim that the evidence he cites *definitively* establishes physicalism, only that it is suggestive of its truth. Perhaps, then, I have set the bar too high in demanding that strong emergence be demonstrably false before endorsing $C_P$? My rejoinder is that I have made no such demand. It is important to bear in mind that functional reduction is a 'case-by-case' process. Property-instances are shown to be realizers of functionalized properties by deducing an instance of the latter from the former. To show that strong emergence is false by functional reduction alone would require deduction of mental property-instances from (let's say) physiological property-instances for all of the possible physiological properties in the putative realization base. If multiple realization is true, then this process could conceivably take forever. However, in 7.2 I argued only that the non-triviality argument requires "a stock of (at least *fairly*) complete non-mental explanations of effects we know to have mental causes" in order to go through. If we have at our disposal a few functional reductions of specific mental property-instances, then I see no problem for the induction – for then it will be true that we have, in the past, successfully given complete explanations of bodily movements in physiological terms. And I for one would have no trouble inducing $C_P$ from such a stock; I would also have no trouble directly inducing physicalism from the same stock. It is not the fact that we do not have a complete set of such explanations that bugs me when I evaluate Melnyk's evidence; it is the fact that we do not have any at all. The putative evidence for $C_P$ fails to rule out strong emergence, not because there isn't *enough* of it, but because it just isn't the right *sort* of evidence.

A second objection suggests itself. What reason have I given to suppose that there *are any* strongly emergent properties? We know, after all, that there are functional properties with physical realizers – properties that are functionally reduced, and which, therefore, we know to supervene on the physical. But strongly emergent

properties, as I have conceived of them, are on rather shakier epistemological grounds. The early British emergentists thought that chemical bonds were emergent; but the reduction of the chemical bond to quantum mechanics put paid to that.[242] What reason is there for thinking that mental properties are emergent? Well I confess, there is none that I can think of at present. But that isn't the point. The causal argument is supposed to discriminate *a priori* between empirical possibilities, and in the process answer metaphysical questions that have yet to be settled scientifically. It matters not one jot that there is no evidence for emergentism; the fact remains that given its *possibility*, there is no current evidence for physicalism either! Notice that I fully endorse the view that there could be evidence for either position: a pattern of explanatory failure suggests emergence, and a pattern of success suggests physicalism. No pattern at all suggests…that more research is needed.

An explanation, even the beginnings of an explanation, of any mental property in physiological (or otherwise non-mental) terms, will count as evidence for $C_P$, and will correspondingly strengthen the non-triviality argument. But it will at the same time count as direct support for physicalism. Reductive explanations go like this: given the laws of physics, and these physical properties, we would expect X to behave like *this*. There is no other way, I have argued, to support the view that X's causal powers are not determined by a strongly emergent property. And this, it goes without saying, is very good evidence that X is nothing over and above the physical (or alternatively, that physical properties are physically sufficient for X). Since we do not have anything like a reduction of mind, I claim that we do not know whether or not $C_P$ is true. This being so, for all we currently know, the causal argument, though valid when combined with our additional arguments of 6.4-6.5, is unsound. As such, until the relevant empirical facts are in, I recommend agnosticism as a metaphysic of mind.

[242] See McLaughlin [1992] for a detailed treatment of this particular case. McLaughlin is of course correct that reduction of chemical bonding showed the latter not to be emergent. However, he goes too far in claiming that this reduction shows that there are *no* emergent properties. Plenty of areas are still up for grabs, including mentality.

# Conclusion

Allow me to summarise the argument of this thesis. The causal argument for physicalism, if sound, establishes that physical events are synchronically sufficient for mental events. It does not equivocate on the sense of 'physical', and relies neither on transmission principles nor causal competition between mental and physical causes for the same effects, in order to go through. The question whether or not the argument is valid, assuming its soundness, amounts to this: are there forms of synchronic sufficiency that are consistent with its premises but inconsistent with physicalism? The causal exclusion argument, if it were sound, would establish not sufficiency but type identity, and there is clearly no question as to whether this latter position is consistent with physicalism or not. But the exclusion argument, although valid, is not sound, as it relies on a theory of causation that is demonstrably false. The causal argument proper relies on no such theory, but is invalid, due to the consistency of weak emergence with its premises. The invalidity of the causal argument is arguably not fatal, for there are independent reasons for rejecting weak emergence. If the argument is sound, then, it forms part of a very strong case for physicalism. Unfortunately, it is not clear whether the causal argument is sound. Strong emergence is not consistent with the premises of the argument, as it involves the existence of properties whose causal powers violate the completeness of physics. This being so, the putative evidence for completeness, if it is good evidence, will be evidence against strong emergence. But the putative evidence for completeness from the current state of science does not tell against strong emergence, and so is not evidence for completeness. However, we can look to strong emergence to see what good evidence for the completeness of physics would look like.

The only way to justify the claim that a property is not strongly emergent is to functionally reduce it to the physical. From this it follows that the only way to justify the completeness of physics, is to functionally reduce any putative strongly emergent properties to the physical. But if a domain of properties is functionally reduced, then we already know that physicalism is true for that domain. Although empirical

evidence of the completeness of physics is possible, it is the sort of evidence the lack of which motivates the causal argument in the first place. Good empirical evidence for the completeness of physics is equally good evidence that physicalism is true. We can make the same point slightly differently, in terms of the completeness of the non-mental. The causal argument for non-mentalism, if sound, will establish that the mental is nothing over and above the non-mental. The completeness of the non-mental is equivalent to the claim that mental properties do not possess any kind of singular novelty. But the only way to establish that mental properties lack singular novelty, is to provide a functional reduction of mental properties to non-mental properties. Put differently, the only way to establish that there is no downwards causation from the mental to the non-mental is to explain the causal powers of mental properties in wholly non-mental terms. But if we have such a reduction, then we already know that the mental is nothing over and above the non-mental, in which case we do not need the causal argument.

The causal argument for physicalism promises a general argument scheme capable of establishing physicalism about any domain of causes that have physical effects. The *prima facie* beauty of the argument consists in the fact that it promises to decide the metaphysics of such causes *a priori*. We now see that the argument fails just where we need it to succeed. Until the relevant functional reductions are in place, we have no good reason to believe that the completeness of physics is true, hence no good reason to believe that the argument is sound. But once these reductions are in place for a given domain, the causes of that domain are transparently physical, their metaphysics determined *a posteriori*. At worst, the causal argument is useful but unsound; at best, it is sound but useless. The question of physicalism belongs to physics.

## BIBLIOGRAPHY


Armstrong, D. (1978). *Universals and Scientific Realism vol.2: A Theory of Universals*. Cambridge: Cambridge University Press.

Armstrong, D. (1986). 'The Nature of Possibility.' *Canadian Journal of Philosophy* 16: 575-94, reprinted in Kim & Sosa (eds.) [1999] pp.184-93.

Baker, L. R. (1993). 'Metaphysics and Mental Content.' Heil and Mele (eds.) [1993] pp.75-95.

Baker, L. R. (1999). 'Unity Without Identity: A New Look at Material Constitution.' *Midwest Studies in Philosophy* 23: 144-65.

Beckermann, A. (1992). 'Supervenience, Emergence, and Reduction.' Beckermann, Flohr & Kim (eds.) [1992] pp.94-118.

Beckermann, A., Flohr, H. and Kim, J. (eds.) (1992). *Emergence or Reduction: Essays on the Prospects of Nonreductive Physicalism*. Berlin: de Gruyter.

Bennett, J. (1988). *Events and Their Names*. Oxford: Clarendon Press.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Block, N. (1997). 'Anti-reductionism slaps back.' *Philosophical Perspectives* 11: 107-32.

Block, N. (2003). 'Do Causal Powers Drain Away?' *Philosophy and Phenomenological Research* 67: 133-50.

Block, N. (ed.) (1980). *Readings in Philosophy of Psychology vol.1*. London: Methuen.

Bohm, D. (1952). 'A Suggested Interpretation of the Quantum Theory in terms of "Hidden" Variables.' Wheeler & Zurek (eds.) [1983] pp.369-96.

Boolos, G. (1984). 'To be is to be a Value of a Variable (or to be Some Values of Some Variables).' *Journal of Philosophy* 81: 430-50.

Broad, C. D. (1925). *The Mind and its Place in Nature*. London: Routledge and Kegan Paul.

Burge, T. (1979). 'Individualism and the Mental.' *Midwest Studies in Philosophy* 4: 73-121.

Burge, T. (1993). 'Mind-Body Causation and Explanatory Practice.' Heil & Mele (eds.) [1993] pp.97-120.

Burke, M. (1992). 'Copper Statues and Pieces of Copper.' *Analysis* 52: 12-17.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Cartwright, N. (1994). 'Fundamentalism vs. the patchwork of laws.' *Proceedings of the Aristotelian Society* 93: 279-92.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Clapp, L. (2001). 'Disjunctive Properties: Multiple Realizations.' *Journal of Philosophy* 98: 111-36.

Collins, J., Hall, E. and Paul, L. (2001). *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Crane, T. (1991). 'Why indeed? Papineau on supervenience.' *Analysis* 51: 32-7.

Crane, T. (1995). 'Mental Causation.' *Proceedings of the Aristotelian Society* supplementary 69: 211-36.

Crane, T. (2001). 'The Significance of Emergence.' Gillett & Loewer (eds.) [2001] pp.207-24.

Crane, T. and Mellor, D. (1990). 'There is no question of Physicalism.' *Mind*, 99: 185-206.

Davidson, D. (1970). 'Mental Events.' Block (ed.) [1980] pp.107-19.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge MA: MIT Press.

Ehring, D. (1999). 'Tropeless in Seattle: the cure for insomnia.' *Analysis* 59: 19-24.

Einstein A., Podolsky B. and Rosen N. (1935). 'Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?' *Physical Review* 47: 777-80.

Elder, C. (2000). 'Mental Causation versus Physical Causation: No Contest.' *Philosophy and Phenomenological Research* 62: 111-27.

Fair, D. (1979). 'Causation and the Flow of Energy.' *Erkenntnis* 14: 219-50.

Fodor, J. (1974). 'Special sciences, or The Disunity of Science as a Working Hypothesis.' *Synthese* 28: 97-115, reprinted in Block (ed.) [1980] pp.120-33.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge MA: MIT Press.

Fodor, J. (1994). *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fodor, J. (1997). 'Special Sciences: Still Autonomous after All These Years.' Horgan (ed.) [1997] pp.149-63.

Funkhouser, E. (2002). 'Three Varieties of Causal Overdetermination.' *Pacific Philosophical Quarterly* 83: 335-51.

Geach, P. (1969). *God and the Soul*. London: Routledge.

Gillett, C. (2002). 'The Dimensions of Realization: A Critique of the Standard View.' *Analysis* 62: 316-23.

Gillett, C. (2003). 'Non-Reductive Realization and Non-Reductive Identity: What Physicalism Does Not Entail.' Walter & Heckmann (eds.) [2003] pp.31-57.

Gillet, C. and Loewer, B. (eds.) (2001). *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

Gillett, C. and Rives, B. (2005). 'The Non-Existence of Determinables: Or, a World of Absolute Determinates as a Default Hypothesis.' *Noûs* 39: 483-504.

Gillet, C. and Witmer, D. G. (2001). 'A "physical" need: Physicalism and the via negativa.' *Analysis* 61: 302-9.

Hall, N. (2000). 'Causation and the Price of Transitivity.' *Journal of Philosophy* 97: 198-222.

Hall, N. (2001). 'Two Concepts of Causation.' Collins, Hall & Paul (eds.) [2001] pp.225–76.

Hardy, L. (1998). 'Spooky Action at a Distance in Quantum Mechanics.' *Contemporary Physics* 39: 419-29.

Hare, R. M. (1963). *The Language of Morals*. Oxford: Clarendon Press.

Heil, J. and Mele, A. (eds.) (1993). *Mental Causation*. Oxford: Clarendon Press.

Hellman, G. and Thompson, F. (1975). 'Physicalism: Ontology, Determination and Reduction.' *Journal of Philosophy* 72: 551-64, reprinted in Kim & Sosa (eds.) [1999] pp.531-39.

Hendel, G. (2001). 'Physicalism, Nothing Buttery and Supervenience.' *Ratio* 14: 252-62.

Hendry, R. F. (1999). 'Molecular models and the question of physicalism.' *Hyle* 5: 143–60.

Hitchcock, C. (2001). 'A Tale of Two Effects.' *Philosophical Review* 110: 361-96.

Horgan, T. (1984). 'Jackson on Physical Information and Qualia.' *Philosophical Quarterly* 34: 147-152.

Horgan, T. (1993). 'From Supervenience to Superdupervenience: Meeting the Demands of a Material World.' *Mind* 102: 555-86.

Horgan, T. (1997). 'Kim on Mental Causation and Causal Exclusion.' Tomberlin (ed.) [1997] pp.165-84.

Horgan, T. (2001). 'Causal Compatibilism and the Exclusion Problem.' *Theoria* 16: 95-116.

Hughes, R. (1992). *The Structure and Interpretation of Quantum Mechanics*. Cambridge, MA: Harvard University Press.

Jackson, F. (1982). 'Epiphenomenal Qualia.' *Philosophical Quarterly* 32: 127-36

Jackson, F. (1994). 'Armchair Metaphysics.' Hawthorne, J and Michael, M (eds.) (1994) *Philosophy in Mind*. Dordrecht: Kluwer pp.23-52.

Jackson, F. and Pettit, P. (1990a). 'Program Explanation: A General Perspective.' *Analysis* 50: 107–17.

Jackson, F. and Pettit, P. (1990b). 'Causation and the philosophy of mind.' *Philosophy and Phenomenological Research* 50: 195-214.

Jackson, F. and Pettit, P. (1992). 'In Defense of Explanatory Ecumenism.' *Economics and Philosophy* 8: 1-22.

Kim, J. (1976). 'Events as Property Exemplifications.' Brand, M and Walton, D (eds.) (1976). *Action Theory*. Dordrecht: Reidel pp.159-77, reprinted in Kim [1993a] pp.33-52.

Kim, J. (1984). 'Epiphenomenal and Supervenient Causation.' *Midwest Studies in Philosophy* 9: 257-70, reprinted in Kim [1993a] pp.92-108.

Kim, J. (1989a). 'Mechanism, Purpose, and Explanatory Exclusion.' *Philosophical Perspectives* 3: 77-108 reprinted in Kim [1993a] pp.237-64.

Kim, J. (1989b). 'The myth of non-reductive materialism.' *Proceedings and Addresses of the American Philosophical Association* 63: 31-47, reprinted in Kim [1993a] pp.265-84.

Kim, J. (1990). 'Supervenience as a Philosophical Concept.' *Metaphilosophy* 21: 1-27, reprinted in Kim & Sosa (eds.) [1999] pp.540-56.

Kim, J. (1992a). 'Multiple Realization and the Metaphysics of Reduction.' *Philosophy and Phenomenological Research* 52: 1–26, reprinted in Kim [1993a] pp.309-35.

Kim, J. (1992b). '"Downward Causation" in Emergentism and Nonreductive Physicalism.' Beckermann, Flohr & Kim (eds.) [1992] pp.119-38.

Kim, J. (1993a). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

Kim, J. (1993b). 'The Non-Reductivist's Troubles with Mental Causation.' Heil & Mele (eds.) [1993] pp.189-210, reprinted in Kim [1993a] pp.336-57.

Kim, J. (1993c). 'Postscripts on mental causation.' Kim [1993a] pp.358-67.

Kim, J. (1998). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, J. (1999a). 'Making sense of emergence.' *Philosophical Studies* 95: 3-36.

Kim, J. (1999b). 'Supervenient Properties and Micro-based Properties: A Reply to Noordhof.' *Proceedings of the Aristotelian Society* 99: 115-18.

Kim, J. (2003). 'Blocking Causal Drainage and Other Maintenance Chores with Mental Causation.' *Philosophy and Phenomenological Research* 67: 151-76.

Kim, J. and Sosa, E. (eds.) (1999). *Metaphysics: An Anthology*. Oxford: Blackwell.

Kistler, M. (1998). 'Reducing Causality to Transmission.' *Erkenntnis* 48: 1-24.

Kistler, M. (2001). 'Causation as transference and responsibility.' Spohn, Ledwig & Esfeld (eds.) [2001] pp.115-33.

Knowles, J. (1999). 'Physicalism, teleology and the miraculous coincidence problem.' *Philosophical Quarterly* 49: 164-81.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lepore, E. and Loewer, B. (1987). 'Mind Matters.' *Journal of Philosophy* 84: 630-42.

Lewis, D. (1966). 'An Argument for the Identity Theory.' *Journal of Philosophy* 63: 17-25.

Lewis, D. (1968). 'Counterpart Theory and Quantified Modal Logic.' *Journal of Philosophy* 65: 113-26.

Lewis, D. (1972). 'Psychophysical and Theoretical Identifications.' *Australasian Journal of Philosophy* 50: 249-58, reprinted in Block (ed.) [1980] pp.207-15.

Lewis, D. (1973). 'Causation.' *Journal of Philosophy* 70: 556-67, reprinted in Lewis [1986d] pp.159-71.

Lewis, D. (1980). 'Mad Pain and Martian Pain.' in Block (ed.) [1980] pp.216-22.

Lewis, D. (1983).'New Work for a Theory of Universals.' *Australasian Journal of Philosophy* 61: 343-77, reprinted in Lewis [1999] pp.8-55.

Lewis, D. (1986a). *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, D. (1986b). 'Against Structural Universals.' *Australasian Journal of Philosophy* 64: 25-46, reprinted in Lewis [1999] pp.78-107.

Lewis, D. (1986c). 'Postscripts to "Causation".' Lewis [1986d] pp.172–213.

Lewis, D. (1986d). *Philosophical Papers: Volume II*. Oxford: Oxford University Press.

Lewis, D. (1992). 'Armstrong on Combinatorial Possibility.' *Australasian Journal of Philosophy* 70: 211-24, reprinted in Lewis [1999] pp.196-214.

Lewis, D. (1999). *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

Lewis, D. (2000). 'Causation as Influence.' *Journal of Philosophy* 97: 182-97.

Lewis, D. and Langton, R (1998). 'Defining Intrinsic.' *Philosophy and Phenomenological Research* 58: 333-45.

Loewer, B. (2001a). 'From Physics to Physicalism.' Gillet & Loewer (eds.) [2001] pp.37-56.

Loewer, B. (2001b). 'Review of "Mind in a Physical World".' *Journal of Philosophy* 98: 315-24.

Lowe, E. J. (1993). 'The Causal Autonomy of the Mental.' *Mind* 102: 629-44.

Lowe, E. J. (2000). 'Causal closure principles and emergentism.' *Philosophy* 75: 571-85.

Lowe, E. J. (2003). 'Physical Causal Closure and the Invisibility of Mental Causation.' in Walter & Heckmann (eds.) [2003] pp.137-54.

Marcus, E. (2001). 'Mental Causation: Unnaturalized but not Unnatural.' *Philosophy and Phenomenological Research* 63: 57-83.

Marras, A. (1998). 'Kim's Principle of Explanatory Exclusion.' *Australasian Journal of Philosophy* 76: 439-51.

Marras, A. (2000). 'Critical Notice of *Mind in a Physical World* by Jaegwon Kim.' *Canadian Journal of Philosophy* 30: 137-60.

Marras, A. (2002). 'Kim On Reduction.' *Erkenntnis* 57: 231-57.

McDermott, M. (1995). 'Redundant Causation.' *British Journal for the Philosophy of Science* 40: 523-44.

McLaughlin, B. (1992). 'The Rise and Fall of British Emergentism.' Beckermann, Flohr & Kim (eds.) [1992] pp.49-93.

Mellor, D. (1995). *The Facts of Causation*. London: Routledge Press.

Melnyk, A. (2003). 'Some Evidence for Physicalism.' Walter & Heckmann (eds.) [2003] pp.155-72.

Menzies, P. (2003). 'The Causal Efficacy of Mental States.' Walter & Heckmann (eds.) [2003] pp.195-223.

Merricks, T. (2001). *Objects and Persons*. Oxford: Clarendon Press.

Millikan, R (1993). *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Montero, B. (2003). 'Varieties of Causal Closure.' Walter & Heckmann (eds.) [2003] pp.173-87.

Morgan, C. L. (1923). *Emergent Evolution*. London: Williams and Norgate.

Noordhof, P. (1997). 'Making the Change: the Functionalist's way.' *British Journal for the Philosophy of Science* 48: 233-50.

Noordhof, P. (1998). 'Do Tropes Resolve the Problem of Mental Causation?' *Philosophical Quarterly* 48: 221-96.

Noordhof, P. (1999a). 'The Overdetermination Argument versus the Cause-and-Essence Principle – No Contest.' *Mind* 108: 367-75.

Noordhof, P. (1999b). 'Micro-based Properties and the Supervenience Argument: A Response to Kim.' *Proceedings of the Aristotelian Society* 99: 109-14.

Noordhof, P. (2003). 'Not Old…But Not That New Either: Explicability, Emergence and the Characterisation of Materialism.' Walter & Heckmann (eds.) [2003] pp.85-108.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge: Harvard University Press.

O'Connor, T. (1994). 'Emergent Properties.' American Philosophical Quarterly 31: 91-104.

O'Connor, T. (2000). 'Causality, Mind and Free Will.' *Philosophical Perspectives* 14: 105-17.

Olson, E. (2001). 'Coinciding Objects and the Indiscernibility Problem.' *Philosophical Quarterly* 51: 337-55.

Papineau, D. (1985). 'Social Facts and Psychological Facts.' Currie, G and Musgrave, A (eds.) (1985) *Popper and the Human Sciences*. Nijhoff: Dordrecht pp.57-71.

Papineau, D. (1989). 'Why supervenience?' *Analysis* 50: 66-71.

Papineau, D. (1991). 'The Reason Why: Response to Crane.' *Analysis* 51: 37-40.

Papineau, D. (1993). *Philosophical Naturalism*. Oxford: Blackwell.

Papineau, D. (1996). 'Many Minds are No Worse than One.' *British Journal for the Philosophy of Science* 47: 233-41.

Papineau, D. (1998) 'Mind the gap.' Tomberlin (ed.) [1998] pp.373-89.

Papineau, D. (2001). 'The Rise of Physicalism.' Gillett & Loewer (eds.) [2001] pp.3-36.

Paul, L. A. (2000). 'Aspect Causation.' *Journal of Philosophy* 97: 235-56.

Paull, C. and Sider, T. (1992). 'In Defense of Global Supervenience.' *Philosophy and Phenomenological Research* 52: 833-54.

Penczek, A. (1997). 'Disjunctive Properties and Causal Efficacy.' *Philosophical Studies* 86: 203-19.

Pereboom, D. (2002). 'Robust Nonreductive Materialism.' *Journal of Philosophy* 99: 499-531.

Putnam, H. (1975a). *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Putnam, H. (1975b). 'Philosophy and our Mental Life.' Putnam [1975a] pp.291-303.

Putnam, H. (1975c). 'The Meaning of "Meaning".' Putnam [1975a] pp.215-71.

Rea, M. (1997). 'Supervenience and Co-Location.' *American Philosophical Quarterly* 34: 367-75.

Rea, M. (1998). 'Sameness Without Identity: an Aristotelian Solution to the Problem of Material Constitution.' *Ratio* 11: 316-28.

Richardson, R. (1979). 'Functionalism and Reductionism.' *Philosophy of Science* 46: 533-58.

Robb, D. (1997). 'The Properties of Mental Causation.' *Philosophical Quarterly* 47: 178–94.

Russell, B. (1914). 'The Relation of Sense-Data to Physics.' *Scientia* 16: 1-27, reprinted in Russell [1917] pp.140-73.

Russell, B. (1917). *Mysticism and Logic*. New York: Doubleday Anchor Books.

Salmon, W. (1984). 'Causal Connections.' Kim and Sosa (eds.) [1999] pp.444-57.

Schaffer, J. (2000). 'Causation by Disconnection.' *Philosophy of Science* 67: 285-300.

Schaffer, J. (2001). 'Causes as Probability-Raisers of Processes.' *Journal of Philosophy* 98: 75-92.

Schaffer, J. (2003). 'Is there a fundamental level?' *Noûs* 37: 498-517.

Segal, G. and Sober, E. (1991). 'The Causal Efficacy of Content.' *Philosophical Studies* 63: 1-30.

Shoemaker, S. (1980). 'Causality and Properties.' van Inwagen, P (ed.) *Time and Cause*. Dordrecht: Reidel pp.109-35, reprinted in Kim and Sosa (eds.) [1999] pp.253-68.

Shoemaker, S. (2001). 'Realization and Mental Causation.' Gillet and Lower (eds.) [2001] pp.74-98.

Sider, T. (2003). 'What's So Bad about Overdetermination?' *Philosophy and Phenomenological Research* 67: 719-26.

Sober, E. (1999). 'The Multiple Realizability Argument Against Reductionism.' *Philosophy of Science* 66: 542-64.

Spohn, W., Ledwig, M. and Esfeld, M. (eds.) (2001). *Current Issues in Causation*. Paderborn: Mentis.

Spurrett, D. and Papineau, D. (1999). 'A Note on the Completeness of "Physics".' *Analysis* 59: 25-9.

Stalnaker, R. (1996). 'Varieties of Supervenience.' *Philosophical Perspectives* 10: 221-42.

Stephan, A. (1997). 'Armchair Arguments Against Emergentism.' *Erkenntnis* 46: 305-14.

Sturgeon, S. (1998). 'Physicalism and overdetermination.' *Mind* 107: 411-32.

Sturgeon, S. (1999). 'Conceptual gaps and odd possibilities.' *Mind* 108: 377-80.

Tomberlin, J. (ed.) (1997). *Philosophical Perspectives, 11: Mind, Causation and World*. Oxford: Blackwell.

Tomberlin, J. (ed.) (1998). *Philosophical Perspectives 12: Language, Mind and Ontology*. Oxford: Blackwell.

van Inwagen, P. (1990). *Material Beings*. New York: Cornell University Press.

Walter, S. and Heckmann, H. (eds.) (2003). *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Exeter: Imprint Academic.

Wheeler, J. and Zurek, W. (eds.) (1983). *Quantum Theory and Measurement*. Princeton: Princeton University Press.

Wigner, E. (1962). 'Remarks on the Mind-Body Question.' Good, I J (ed.) (1962) *The Scientist Speculates: An Anthology of Partly-Baked Ideas*. London: Heinemann pp.284-302, reprinted in Wheeler & Zurek (eds.) [1983] pp.168-81.

Wilson, J.(2005). 'Supervenience-based Formulations of Physicalism.' *Noûs* 39: 426-59.

Witmer, D. G. (2000). 'Locating the Overdetermination Problem.' *British Journal for the Philosophy of Science* 51: 273–86.

Witmer, D. G. (2001). 'Sufficiency Claims and Physicalism: A Formulation.' Gillett & Loewer (eds.) [2001] pp.57-73.

Witmer, D. G. (2003). 'Functionalism and Causal Exclusion.' *Pacific Philosophical Quarterly* 84: 198-214.

Yablo, S. (1992). 'Mental Causation.' *Philosophical Review* 101: 245-80.

Yablo, S. (2002). 'De Facto Dependence.' *Journal of Philosophy* 99: 130-48.